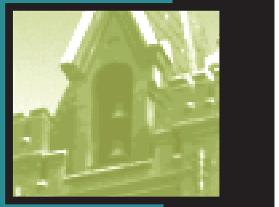
Analytic Issues in the Assessment of Student Achievement

Analytic Issues in the Assessment of Student Achievement



Proceedings from a Research Seminar Jointly Sponsored by

National Center for Education Statistics
National Institute on Student Achievement, Curriculum and Assessment
and
RAND

U.S. Department of Education

NCES 2000-050

The Charles Sumner School

More than any other school founded after the Civil War, the Charles Sumner School served as the cornerstone for the development of educational opportunities for black citizens in the District of Columbia. The significance assigned to its design and construction was indicated by the selection of Adolph Cluss as architect for the new building. In 1869, Cluss had completed the Benjamin Franklin School; in 1872, he completed Sumner School; and in 1873, he won a medal for “Progress in Education and School Architecture” for the City of Washington at the International Exposition in Vienna, Austria.

Dedicated on September 2, 1872, the new school was named in honor of United States Senator Charles Sumner of Massachusetts, who ranked alongside Abraham Lincoln and Thaddeus Stevens in leading the struggle for abolition, integration, and nondiscrimination. Upon opening, the Sumner building housed eight primary and grammar schools, as well as the executive offices of the Superintendent and Board of Trustees of the Colored Schools of Washington and Georgetown. The building also housed a secondary school, with the first high school graduation for black students held in 1877. The school also offered health clinics and adult education night classes.

A recipient of major national and local awards for excellence in restoration, Sumner School currently houses a museum, an archival library, and other cultural programs that focus on the history of public education in the District of Columbia.

Note: Cover photographs of the Charles Sumner School by David W. Grissmer

Analytic Issues in the Assessment of Student Achievement

David W. Grissmer
and
J. Michael Ross
Editors

Proceedings from a research seminar jointly sponsored by
National Center for Education Statistics
National Institute on Student Achievement, Curriculum, and Assessment
RAND

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006-5574

June 2000

The NCES World Wide Web Home Page is http://nces.ed.gov.

Suggested Citation: U.S. Department of Education. National Center for Education Statistics. *Analytic Issues in the Assessment of Student Achievement,* NCES 2000–050 by David W. Grissmer and J. Michael Ross. Washingon, DC: 2000.

For ordering information on this report, write:
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

Contact:
David Grissmer
(703) 413–1100 ext. 5310
J. Michael Ross
(202) 502–7443

Or call toll free 1–877–4ED–Pubs.

Dedication

Dedicated to David W. Stevenson (1951–1998)

Senior Advisor to the Acting Deputy Secretary of Education, 1993–98

This book is dedicated to the memory of David W. Stevenson. His understanding of the interplay between basic research and education policy facilitated the development of this research seminar. From his early days in the sociology program at Yale, David began to develop a discipline-specific understanding of the structural factors mediating social change. As he became more involved in controversial policy issues, he saw the necessity for more definitive empirical evidence in their resolution. In the continual efforts of the research and policy communities, David's perspective will continue to enrich conversations about the direction of and appropriate methodologies for education reform. We acknowledge, with this dedication, his memorable accomplishments and our appreciation for his influence on this research seminar.

Table of Contents

Foreword ix
Peggy G. Carr

Acknowledgments xi
David W. Grissmer and J. Michael Ross, Editors

Introduction: Toward Heuristic Models of Student Outcomes and More Effective Policy Interventions........... xv
C. Kent McGuire

SECTION I.
USING EXPERIMENTS AND STATE-LEVEL DATA TO ASSESS STUDENT ACHIEVEMENT

Synthesizing Results from the NAEP Trial State Assessment 3
Stephen W. Raudenbush

Moving Educational Research Toward Scientific Consensus 43
David W. Grissmer and Ann Flanagan

Response: Guidance for Future Directions in Improving the Use of NAEP Data 91
Sylvia T. Johnson

Section II. Using Longitudinal Data to Assess Student Achievement

Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data 103
Meredith Phillips

Certification Test Scores, Teacher Quality, and Student Achievement 133
Ronald F. Ferguson with Jordana Brown

Response: Two Studies of Academic Achievement 157
Robert M. Hauser

Section III. Relating Family and Schooling Characteristics to Academic Achievement

Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88..... 169
Dominic J. Brewer and Dan D. Goldhaber

School-level Correlates of Reading and Mathematics Achievement in Public Schools 189
Donald McLaughlin and Gili Drori

Response: Opportunities for Design Changes 237
Valerie E. Lee

SECTION IV.
POLICY PERSPECTIVES AND CONCLUDING COMMENTARY

Assessment Trends in a Contemporary Policy Context 249
Marshall S. Smith

Educational Research and Educational Policy: An Historical Perspective 279
Christopher Jencks

Improving Research and Data Collection on Student Achievement 299
Brenda J. Turnbull, David W. Grissmer, and J. Michael Ross

SECTION V.
APPENDIX

Seminar Attendees 319

Foreword

Peggy G. Carr
Associate Commissioner
Assessment Division
National Center for Education Statistics

In November 1998, a group of outstanding researchers and scholars gathered at the Charles Sumner School in Washington, DC to explore methodological issues related to the measurement of student achievement. Within this broad topic, the research seminar also focused more specifically on the sharing of perspectives related to the black-white test score gap. This sharing enabled the participants to compare their analyses and findings and to recommend improvements in data collection and analysis to the National Center for Education Statistics (NCES). Thus, eventually this collegial exchange promises to improve the utility of NCES data sets for policymakers in their efforts to ensure both excellence and equity in American education.

Seeking deeper explanations of the test score gap is a critical first step in the process of assessing student achievement more accurately. Toward that end, the seminar demonstrated the need for NCES to pursue more aggressively the development of concepts and methodologies that allow independent analysts to unravel the causes of such gaps. Such an “unraveling” requires closer examination of the complex interrelationships among resource factors, home and schooling influences, family configurations, and achievement outcomes. Further, NCES needs to place both cross-sectional and longitudinal data in a broader framework and to explicate our findings within diverse social contexts in richer detail.

The work of the Assessment Division in NCES, in particular, will benefit from the development of more explicit constructs that allow better comparisons of achievement results without the confounding interpretations that typically characterize conventional statistical presentations. For example, when achievement discrepancies between blacks and whites reveal different patterns in the northern states as compared to southern states, what type of analysis can we conduct that would enlighten our understanding of these historical and contemporary differences?

This first seminar has reminded us of the value of having researchers, scholars, and practitioners come together to advance knowledge in the field of achievement research and assessment. The collaboration of the sponsoring agencies—NCES, RAND, and the Office of Educational Research and Improvement (OERI) and its Achievement Institute—with their different missions, exemplifies the desire to integrate discipline-based perspectives toward common education reform goals. OERI and NCES acknowledge ongoing opportunities to sponsor a series of research seminars in order to ensure continued progress toward improving education policies and practices on behalf of our children and youth.

Seeking to engage a broader audience in this collegial exchange, NCES has prepared this volume containing the papers originally presented at the Charles Sumner School. The exchange of ideas among researchers and policymakers remains important to NCES. Still, this publication does not necessarily reflect the views of NCES or the policies of the U.S. Department of Education. Rather, the papers included here represent the views of their respective authors alone.

Acknowledgments

David W. Grissmer and J. Michael Ross Editors

The idea of a “research seminar” where academic researchers could share their current research findings with their federal counterparts took shape initially in early 1997. Ongoing discussions about the potential benefits of collaboration among the National Center for Education Statistics (NCES), RAND, and the National Institute on Student Achievement, Curriculum, and Assessment (NISACA) gave rise over the next year to a conceptual structure. A number of common interests were identified in the research and policy communities: periodic updates on complex survey designs and multilevel types of analysis. We went on to consider also our broader purposes: providing the direction to research that will inform policy developments in education, generating wider awareness of education research, and stimulating the development of better educational theory.

Within NCES, new forms of collaborative exchange were discussed. The one-day seminar received early support from Gary Phillips and Peggy Carr of the Assessment Division. Sharif Shakarani, then of the Assessment Division, helped to focus seminar offerings on NCES issues in data collection and analysis and fostered further collaboration by endorsing the participation of the different divisions in such a conference. Their understanding of the relevance of research updates shaped the concepts under discussion toward NCES needs. We are grateful, too, for Peggy’s strong and continuous advocacy and her financial support for the seminar. We appreciate also the substantive support offered by Holly Spurlock of the Assessment Division, whose careful and competent assistance throughout the process proved invaluable to the eventual success of the seminar. During this time, Daniel Kasprzyk, Director of the Schools and Staffing Program of NCES, also provided critical financial and moral support, and we remain grateful for his early commitment. The emerging plans for the seminar received support from Pascal D. Forgione, then Commissioner of NCES, whose sentiments were always directed toward providing the best research possible in the interests of assisting policymakers to improve education.

As the planning progressed, Joseph Conaty of NISACA provided ongoing insights in organizational support through his contacts with the academic research community. For the critical collaborations he contributed to this endeavor, we express our continuing appreciation. Further, we acknowledge the contributions of Marian Robinson, then an intern in Joseph's office, now at the Graduate School of Education, Harvard University, who smoothly executed numerous details of planning for the seminar.

We would also like to thank Marilyn McMillen, Chief Statistician for NCES, for broadening the base of participation in the seminar through the provision of special funding to cover the travel expenses of graduate students.

The role of the Education Statistics Services Institute (ESSI) in furthering the broad research and development purposes of the seminar is also very much appreciated. ESSI's ability to facilitate "making the seminar happen" made it possible for us to extend collaboration and consultation among all the participating groups. Another colleague whose support was critically important in the early stages of development is John Mullens, now of Mathematica Policy Research, who worked on the project under the auspices of ESSI. John offered substantive contributions to discussions about the importance and structure of the seminar, and then cheerfully took the lead in facilitating arrangements among all the parties. Later, he played an important role in ensuring that the early drafts of the solicited papers arrived in time for review before they were distributed to seminar participants. The benefits of the seminar were enhanced by John's grasp of the issues in research and policy and his facilitative skill.

Our appreciation for managing critical details extends to Bridget Bradley, then a consultant with Policy Studies Associates and later Policy Analyst in the Office of the Deputy Secretary of Education, who offered invaluable logistical support to our efforts to plan the seminar. Her gracious manner complemented her careful attention to making and monitoring arrangements, and we thank her sincerely for her efforts.

We extend *very* special thanks to the organizing committee that had major responsibilities for planning and staging the seminar, as follows: Peggy Carr, Holly Spurlock, and Daniel Kasprzyk, as well as John Mullens. We very much appreciate the committee's efforts, through the seemingly endless meetings, messages, and phone calls. Further, this committee, along with Brenda Turnbull of Policy Studies Associates and Martin Orland of the Early Childhood and International Crosscutting Studies Division (ECICSD), also participated in the detailed planning for the publication of the proceedings, and we are indebted to them for their useful suggestions regarding major decisions about this book. The benefits of their efforts on behalf of the seminar should be seen for years to come, as NCES endeavors to ensure continuous improvements in data quality and analytical methods.

On November 9, 1998 at the Charles Sumner School in Washington, DC, the seminar took place with approximately 100 participants in attendance. Titled “Analytic Issues in the Assessment of Student Achievement,” the research seminar was jointly sponsored by NCES; the National Institute on Student Achievement, Curriculum, and Assessment; and RAND, as we had planned for so many months. The beautiful setting, the quality of the papers and the commentary, and the collaborative and collegial nature of the day’s deliberations were the fruition of the long process of preparation.

With appreciation, we acknowledge the “silent” reviewers of the early drafts of the solicited research papers. Their early reviews increased the usefulness and applicability of the presentations and papers. These reviewers, in addition to the editors, were Martin Orland, John Ralph, Dan Kasprzyk, Peggy Carr, Joseph Conaty, and Holly Spurlock. Their work, though behind the scenes, was an important contribution to the substance of the seminar, and we appreciate their assiduous reviews.

Subsequently, the papers were forwarded to the colleagues who had agreed to serve as discussants for the seminar. Sylvia Johnson (Professor of Education at Howard University), Robert M. Hauser (Professor of Sociology at the University of Wisconsin-Madison), and Valerie E. Lee (Professor of Education at the University of Michigan) undertook the task of reviewing each pair of solicited research papers representing the methodological and conceptual strands of the seminar, seen here in Sections I, II, and III. Their comments enabled the authors of the solicited papers to make further improvements in their works before the seminar; then the discussants prepared their public responses for the presentations made during the seminar. We remain grateful for their dedication to this time-consuming task that benefited all seminar participants.

Similarly, we offer our appreciation to Marshall S. Smith and Christopher Jencks, whose presentations lifted our attention from such narrow topics as sampling design and dataset linkages to take a broader look at the effects of past analytical methods upon social scientists' understanding of achievement disparities and to share insights into how those understandings have played a role in the development of new education policies. Smith and Jencks, each in his own way and from his own perspective, explained the vagaries of education research since "the Coleman report" and went on to describe the usefulness of better data collection and analysis and of better theories and models.

Further, we acknowledge with appreciation the assistance of Joseph Conaty, John Ralph, and Martin Orland as moderators for the discussions during the seminar, as well as the participation of the seminar attendees (listed in the appendix), whose comments enriched the discussions and, therefore, the overall outcomes of the seminar.

Following the event, we made the decision to edit the proceedings for publication, recognizing the far-reaching implications of the discussions for NCES and desiring to extend the insights to a broader audience. Even more ambitious were our later decisions to include the Introduction and the fourth section, Policy Perspectives and Concluding Commentary. It was fortunate that Anne Meek of ESSI was available for the tasks that these decisions required. As a professional editor working closely with us, Anne ensured both the completion of the book and its internal coherence. We acknowledge with appreciation her grace and her sense of humor throughout the process of preparation.

In the preparation of this book, special thanks are due to Ron Miller of RAND for the design of the cover of the book (which incorporates a photograph by David Grissmer). We also acknowledge the assistance of staff at ESSI who prepared the proceedings for publication, as follows: Allison Arnold, Mariel Escudero, Anne Kotchek, Qiwu Liu, Jennie Romolo, and Jennifer Thompson. We thank them for their attention to detail and their technical skills, which have greatly improved this book for use by researchers, policymakers, and educators.

The persons named here have provided varied kinds and levels of support for the seminar and for the production of this book, and we are pleased to acknowledge our debt to each of them. However, the final responsibility for this publication rests with us, and any remaining deficiencies are solely our responsibility.

Introduction

Toward Heuristic Models of Student Outcomes and More Effective Policy Interventions

C. Kent McGuire
Assistant Secretary
Office of Educational Research and Improvement

In November 1998, in the research seminar commemorated here in this volume, a diverse community of scholars and researchers paused amidst their heavy schedules to turn their attention to a questioning of their methods of conducting empirical inquiries. Taking stock of a body of work is, of course, commendable for a professional group. It is always instructive to learn from one another and to consider how to better our efforts; and this seminar provided ample opportunity for such learning and consideration along several dimensions.

The seminar, however, went beyond the normal technical matters that education researchers typically discuss on such occasions. Rather, the gathering also shed light on research and policy issues, especially the continuing efforts to improve the performance of American education, to enhance greater educational equality of opportunity, and to understand the sources of continuing race-ethnicity achievement discrepancies. These larger purposes are, after all, the reasons we collect and analyze data in the first place and the reasons we search for improvement in our methods of data collection and analysis.

That the deliberations took place at the Charles Sumner School was especially appropriate for the Office of Educational Research and Improvement (OERI). Sumner School, now restored and an architectural treasure of great beauty, has long served as an important symbol of minority education. In this setting, we were surrounded by a particularly fitting sense of history for this discussion of both the means for measuring student achievement and the reasons for doing so.

The deliberations were enriched by multiple disciplinary perspectives. The research seminar included sociologists, economists, and education researchers, both new and more established researchers, and federal policymakers, all of whom shared their insights with each other. That is, researchers from different disciplines and methodological backgrounds commented on each other's analyses and listened to each other's recommendations, and federal policymakers provided their perspectives on the role of research and the important questions that must be addressed. In short, the seminar provided an enlightening forum for the exchange of perspectives and research findings, as participants contributed their particular expertise to discussions about the measurement of achievement and the contribution of education research to the improvement of schooling.

Of particular importance are some new insights in the understanding of racial and ethnic differences in student achievement. Such differences were first brought to our attention nearly 30 years ago by "the Coleman report," when the nation began to move *equality of educational opportunity* to its enduring place on the nation's agenda. Since then, we have come to understand much more about the variables associated with both high and low achievement—not nearly as much we would like to know but certainly more than we once knew. And OERI has always hoped to play a pivotal role in the empirical examination of these questions.

Over the past 10 to 20 years, the federal government has been improving its data collections, and a wide array of analyses continue to be conducted to move our understanding beyond Coleman's findings. These continued adjustments and processes have helped us to understand the complexity of what we are trying to measure and what we are trying to change. A brief synthesis of the papers solicited for this seminar will serve to illustrate the details of different data sets and, at the same time, help us to understand the systemic obstacles to changes in educational policies.

The papers are organized under three major divisions: (1) Using Experiments and State-level Data to Assess Student Achievement, (2) Using Longitudinal Data to Assess Student Achievement, and (3) Relating Family and Schooling Characteristics to Academic Achievement. The last major division, (4) Policy Perspectives and Concluding Commentary, presents important observations about research methodology and funding and the connections between research and policy, both with a retrospective view and a view toward the future.

Using Experiments and State-level Data to Assess Student Achievement

In the first essay, Stephen Raudenbush characterizes the state proficiency means from the Trial State Assessment of the National Assessment of Educational Progress (NAEP) as “difficult to interpret and misleading.” It is their multidimensionality that makes proficiency scores difficult to interpret: they may look simple at first glance, but actually they reflect many factors—student demographics, school organization and processes, and state policy influences. Raudenbush discusses his multilevel analyses that compare states on their provision of student resources for learning. Not surprisingly, he finds that socially disadvantaged students and ethnic minority students (particularly African American, Hispanic American, and Native American) are significantly less likely than other students to have access to advanced course-taking opportunities, favorable school climates, highly educated teachers, and cognitively stimulating classrooms. He also finds substantial variation across states in the extent of inequality in access to such resources. Such findings point, as he said, toward “sharply defined policy debates concerning ways to improve education.”

Grissmer and Flanagan speak from a different but equally illuminating perspective. Their major focus, fueled by concerns about inconsistency in research results, is the lack of consensus across the broad and multidisciplinary research communities in educational research. In many respects, of course, this lack of consensus has been inevitable, given the different research perspectives; the varied points of view expressed by researchers, policymakers, and practitioners; and the inherent complexity of education. Grissmer and Flanagan believe, therefore, that improvements in data collection and statistical methodologies, by themselves, are not sufficient to bring about the kind of consensus needed to effectively guide educational policies. Thirty years of research with nonexperimental data have led to almost no consensus on important policy issues, such as the effects of educational resources and educational policies on children and the impact of resources on educational outcomes. Further, they propose to guide the process of creating consensus through the development of a strategic plan, which would enable experimentation and data collection to provide the quality of data necessary for theory-building and also improve the specifications of models used in nonexperimental analysis.

Grissmer and Flanagan therefore recommend three approaches likely to lead to consensus: increasing experimentation, building theories of educational process, and improving nonexperimental analysis. They suggest that experiments have two main purposes: they provide the closest-to-causal explanations possible in the social sciences, and they help to validate model specifications for nonexperimental data. They present detailed discussions of important policy issues and the findings of research, including critical analyses of the “money doesn’t matter” issue and the issue of the effects of resources on achievement, with examples from the many ways researchers have addressed these questions over the years. They also provide insight into such efforts as the Tennessee class size experiment, the use of NAEP scores and SAT scores, and new methods of analyzing education expenditures.

In addition to making some methodological recommendations, Grissmer and Flanagan explain the process of theory-building cogently and clearly. To advance theory-building, they advocate linking the disparate and isolated fields of research in education, for example, linking the micro-research on time, repetition, and review with the research on specific instructional techniques, homework, tutoring, class size, and teacher characteristics. Further, to enhance the development of modeling assumptions, they recommend linking the research on physical, emotional, and social development, differences in children, delays in development, and resiliency factors. Their suggestions for improvements encompass the need for experiments, improvements in NAEP data such as collecting additional variables from children, and supplemental data from teachers, among other things. All in all, their paper offers timely and thought-provoking views about the research community’s next steps in improving theories of education and models of research, so that eventually the nation can indeed achieve its desired goals in education.

Using Longitudinal Data to Assess Student Achievement

Next, Meredith Phillips offers a number of convincing and far-reaching observations about improving methods of data collection and analysis, especially in efforts to understand ethnic differences in academic performance. Perhaps most relevant is her observation, echoed by other presenters, that we must study ethnic differences explicitly despite their political sensitivity. She explains that socioeconomic factors do not overlap with ethnicity as much as researchers have traditionally assumed. Ethnic differences in learning vary between the school year and the summer; therefore, the importance of collecting data in both spring and fall of each school year should be a major point of empirical queries. Further, since the test score gap widens more during elementary school than during high school, and children's test scores appear less stable during elementary school than during high school, Phillips also calls for focusing more surveys on elementary students rather than on high school students. Of particular interest is her assertion that we have learned little about ethnic differences because researchers have not adequately studied education outside of the formal institution of schooling. Measuring the cognitive skills of infants and toddlers prior to their entry into school could help to clarify ethnic differences in family influences on achievement. Phillips concludes by reminding us that "it is not logically necessary to understand the causes of a social problem before intervening successfully to fix it." To those who bear responsibility for the improvement of American education, this reminder is somewhat comforting, in view of the breadth and depth of recommendations made by this network of researchers and scholars.

Ferguson and Brown then discuss the relationship of teacher quality to student achievement, in particular, the relationship of teachers' certification test scores to students' test scores. The evidence they have assembled suggests that the black-white test score gap among students reflects a similar test score gap among teachers. From several studies, they cite findings suggesting that "teachers' test scores do help in predicting their students' achievement." For example, scores on the Texas Examination of Current Administrators and Teachers (TECAT) turned out to be strong predictors of higher student reading and math scores in school districts across the state. Ferguson and Brown explicitly make the point that ensuring well-qualified teachers in districts where minority students are heavily represented is "part of the unfinished business of equalizing educational opportunity." In Alabama, certification testing reduced entry into teaching by candidates with weak basic skills and consequently narrowed the skills gap between new black and white teachers. Since the rejected candidates would probably have taught disproportionately in black districts, Ferguson and Brown suggest that the policy of initial certification testing is probably helping to narrow the test score gap between black and white students in Alabama. Predictive validity has not yet been used as a criterion for validating such exams; still, Ferguson and Brown contend that policymakers can safely assume a positive causal relationship between students' and teachers' scores.

Relating Family and Schooling Characteristics to Academic Achievement

Brewer and Goldhaber offer additional insights into the relationship of student achievement and teacher qualifications, based on their analyses of data from the National Education Longitudinal Study of 1988 (NELS:88). Their linking of student-teacher-class elements in NELS:88 permitted these researchers to investigate the effects of specific class size, teacher characteristics, and peer effects on student achievement, through the use of multivariate statistical models. The NELS:88 data enabled the researchers to link students to their particular teachers and specific courses. In their analyses, they find that subject-specific teacher background in math and science is positively related to student achievement in those subjects, as compared to teachers with no advanced degrees or with degrees in non-math subjects. They did not see this pattern repeated in English and history. Nor did they find positive effects on achievement associated with teacher certification or years of teaching experience.

While encouraged by the recent improvements in data collection exemplified by NELS:88, Brewer and Goldhaber make pertinent recommendations for future data collections. Seeing the link between students and teachers as critical, they strongly recommend that such links not only be maintained, but also strengthened by the collection of additional data about teachers' backgrounds. Specifically, they suggest the addition of teacher test scores, the years that teachers obtained their licenses, and the states where they were licensed. Such data would be quite useful now and in the future, since policymakers in many states have recently overhauled or are considering changing licensure and/or teacher preparation requirements.

Brewer and Goldhaber point out that items relating to student, parent, and teacher beliefs, attitudes, and feelings could be omitted from data collections, since policymakers can only indirectly affect these. Further, they raise the questions of de-emphasizing the collection of nationally representative samples or of sampling fewer schools with more data on students and classes in a smaller number of schools. Brewer and Goldhaber are seeking the data quality necessary for the use of multivariate statistical models, because researchers find such models most persuasive in tackling important policy questions. Brewer and Goldhaber clearly state their belief that the “ultimate reason to collect data is to influence public policy in a positive way,” a perspective that supports the continued improvement of data collection and methods of analysis.

Finally, in their investigation into school-level correlates of student achievement, McLaughlin and Drori report linking three sources of data: (1) data from the Schools and Staffing Survey (SASS) regarding such school and background factors as school size, class size, normative cohesion, teacher influence, student behavioral climate, teacher qualifications, and the like; (2) student achievement data from statewide assessments; and (3) data from the 1994 State NAEP fourth grade reading assessment in public schools. These researchers constructed a set of 18 composites of data on student background, organizational aspects, teachers’ qualification, and school climate perceptions, then merged them with school reading and mathematics mean scores. McLaughlin and Drori analyzed the relationships of various school organizational factors to student achievement, hoping to elicit evidence on the correlations between school reform policies and achievement. An important finding is that reading scores were higher in schools with smaller class sizes. This finding was consistent across grade levels. Another interesting finding is that middle and secondary schools in which teachers perceive that they have more than average control over classroom practices and influence on school policies tend to be schools in which mathematics scores are higher.

Perhaps more exciting than their findings, however, is the methodology McLaughlin and Drori employed and its potential for identifying effective school policies. Teasing out the correlates of student achievement through such linkages of databases is a promising venue for researchers and policymakers alike, especially since a number of states are turning to reforms that establish consequences for schools based on their gains in achievement over years.

Policy Perspectives and Concluding Commentary

Midway through the seminar, Marshall S. Smith engaged seminar participants in a retrospective look at past policy efforts to monitor and mitigate the discrepancies in black-white achievement scores. In his paper, he discusses possible explanations for the status of the gap at various points in time and concludes by reviewing current policy directions that promise further improvements in student achievement and recommending increased attention to experimental field trials.

Smith describes the reductions in the black-white achievement gap from 1971 through 1988, as seen in data from NAEP assessments, referring to a paper that he and Jennifer O’Day published in 1991, which reviewed policy initiatives and changes in student achievement 25 years after the Coleman report. Smith, who was at that time dean of the graduate school of education at Stanford University, pointed out in his presentation that these reductions reflected consistent and substantial increases in black scores and almost no change in white scores. In less than 20 years, the reduction in the achievement gap between black and white students was 33-50 percent in reading and 25–40 percent in mathematics, according to NAEP data.

Smith summarizes several tentative explanations for this reduction in the gap, which occurred between 1971 and 1988, which he and O’Day had first discussed in their paper. They had recognized, first, the large decrease in the percentage of black children living in poverty: from 65 percent in 1960 to 42 percent in 1980. Another highly plausible explanation was that preschool attendance increased substantially for low-income children. Further, Smith notes, the educational quality of schools for black students was dramatically enhanced with the dismantling of the old dual school system. In addition, the effects of Title I—while difficult to assess by numbers alone—included an increase in educational resources in schools, lower class sizes, and an emphasis on the basics of reading and mathematics. And, as Smith reiterated during the seminar, Title I also served to focus national attention on the needs of low-income students, many of whom were African American.

Smith reminds seminar participants that he considered the basic skills movement an influence in reducing the achievement gap at the secondary level during this period. After all, by the mid-1980s over 33 states had required students to pass a minimum competency test as a criterion for graduation. The resulting instructional emphasis on basic skills, combined with the “high stakes” tests, produced the focus and coherence in the curriculum needed for improving student achievement.

Smith goes on to speculate that, by 1990, the effects of the factors identified by him and O’Day had begun to diminish in their influence and that, therefore, the gap between black and white students’ test scores was no longer continuing to narrow. Thus, the current task for policymakers has become to identify and implement policy ideas that promise to continue the process of reducing the gap initiated in prior decades. This task means thinking hard about, and also building upon, the interventions that brought about the earlier improvements in achievement.

Smith describes three major objectives at the federal level designed to support efforts to improve education in general and also to reduce the gap. The first is to create overall conditions as stable and livable as possible for all families with children. Smith cites, as efforts toward this objective, recent sustained economic growth and specific policies such as the Earned Income Tax Credit and the Children's Health Insurance Plan. The second objective is to expand educationally rich opportunities for all students beyond typical school schedules. As specific examples Smith lists the development of education standards for the Head Start curriculum, the expansion of Head Start enrollment, and increased services through the 21st Century After-School Program. The third is to encourage state and local standards-based reforms. Toward this end, federal programs such as Title I and Goals 2000 have been aligned to support the state reforms.

Standards-based reform, considered one of Smith's major contributions to education policy, in effect extends the basic skills movement to a much broader scope, with *all* children expected to attain the higher content and performance standards, not just basic skills. Even at such an early date as this, it is worth examining the promise of such reforms by looking at outcomes within the states. What have been the test score results in states with focused and coherent strategies in their standards-based reforms? Using NAEP data, Smith finds encouraging results in those states—especially North Carolina and Texas—with relatively challenging standards, curriculum-aligned tests, accountability provisions, extensive teacher training, and special efforts on behalf of low-scoring students. It is apparent that, for whatever reasons, some states are doing very well in their efforts to improve student outcomes, while others are not. Therefore, policymakers are obliged to consider very carefully the evidence about interventions that promise to lead to improved student performance.

Moving to a prospective view, and building a case for increasing experimental efforts, Smith cites the strength and authority of such studies as the Tennessee class size study and those on early reading acquisition at the National Institute of Child Health and Human Development (NICHD). He identifies several areas where policy development could well be more adequately informed through such studies; for example, methods of incorporating technology into classrooms, the effects of summer school, and replications of the NICHD studies. Smith argues eloquently for increasing the use of experimental field trials in education research and suggests that a list of recommendations for consideration for the research agenda at the Department of Education might come from the seminar.

Indeed, as Christopher Jencks pointed out in his presentation, for those who believe that educational policy should be based upon a more solid evidentiary structure, the current shortage of any type of randomized field trials in education policy represents perhaps the greatest challenge facing education policymakers and researchers alike. More pointedly, of course, OERI faces this challenge in designing a course for its own research agenda. According to Jencks, a major advantage of experimental studies is that the more persistent and difficult policy questions can be answered more definitively by the inclusion of randomization procedures at the school and classroom levels. These questions cannot be answered by improved data collections, more complex surveys, or more refined statistical methods alone. Critical policy questions such as the debate over ability grouping can be intensely controversial; and to resolve such questions by randomized field trials would still entail some unavoidable political fallout, no matter how definitive the findings.

Then, too, Jencks notes that the idea of randomized trials is rarely accepted within the field of education research. There are a number of practical obstacles to utilizing experimental methods: they inevitably change established school routines, since they necessarily include randomization of students or teachers to different schools or classes. It might be possible to convince educators that such procedures would constitute a small price to pay, given the very useful information to be gained, if only the researchers themselves strongly supported experimental studies. Jencks notes, however, that most education researchers are typically unenthusiastic about randomized experiments. In fact, he contends that most researchers now have limited knowledge of classic experimental studies.

Still, Jencks insists that the advantages of randomized field trials to policymakers are large and attractive. The first advantage lies in the knowledge to be gained from wider use of experimental methods; the second, in the clarity of understanding that results from these intuitively obvious methods. A legislator or a school board member, for example, can follow the logic of the Tennessee class size experiment, understand how the results were evaluated, and see why the results are consistent with what researchers say they mean. Nevertheless, Jencks is not suggesting that we abandon descriptive types of research proposals. On the contrary, surveys and experiments complement one another, each yielding valuable results necessary for providing the data necessary for policymaking. But the present dearth of experiments sounds a warning to OERI and highlights an imperative need for the next few years.

Indeed, with such different perspectives and challenging viewpoints brought to bear on a single topic, many possible directions were identified for the future work of NCES and OERI. Throughout the seminar, presenters and participants were persuasive in their descriptions of the necessity of complementing longitudinal survey data with data collected in the classical research design tradition such as the Tennessee class size experiment. Their praise for renewed consideration of experiments made this issue the predominant theme of the seminar, and one with far-reaching implications for the sponsors of the event.

Taking stock of our empirical methods—more or less the primary reason for organizing the seminar—yielded a second theme in the comments from presenters and participants. This theme was seen in the abundance of proposals for improvements in the design and analysis of data collections, including ways of making longitudinal studies more elaborate; suggestions about the addition or deletion of certain types of items on surveys; sampling more students per teacher; collecting longitudinal data more frequently; and gathering more measures of teacher quality. Implicit in many of the recommendations is the idea of more critical evaluation of the utility of variables and methods in all NCES surveys, whether longitudinal or cross-sectional, in order to design better surveys in the future. These suggestions translate into serious considerations for OERI and NCES as they move forward with new assessments of student achievement, as well as with all other surveys and analyses.

Last but by no means least, seminar participants emphasized the importance of communication among the different research disciplines. They referred specifically to the power of experiments to communicate effectively with policymakers and other researchers. They expressed appreciation for the seminar as a good example of such communication and recommended more such opportunities. The value of the seminar can easily be seen in the broad, data-based dialogue among researchers about the choices facing NCES and OERI and presented in this book. Suggestions were made to open the door to new partnerships among federal, state, and private researchers and to establish connections between state-based researchers and federal researchers. Interestingly enough, repeated references to the benefits to be gained from openness to a variety of audiences constituted a sub-theme of the seminar. Communication is, after all, an essential component of building consensus among researchers, scholars, and policymakers.

In short, the exchanges of this seminar promise researchers and policymakers alike that racial and ethnic differences in achievement can be explored more effectively than at present, that schools can continue to move toward equality of educational opportunity, and that progress toward the improvement of American education requires our continued communication, collaboration, and commitment. It is now our task to translate our knowledge into improved policies and practices in education for the benefit of our children and our nation.

SECTION I.
USING EXPERIMENTS AND STATE-LEVEL DATA TO ASSESS STUDENT ACHIEVEMENT

Synthesizing Results from the NAEP Trial State Assessment 3
Stephen W. Raudenbush

Moving Educational Research toward Scientific Consensus 43
David W. Grissmer and Ann Flanagan

Response: Guidance for Future Directions in Improving the Use of NAEP Data 91
Sylvia T. Johnson

Synthesizing Results from the NAEP Trial State Assessment

Stephen W. Raudenbush
School of Education and Survey Research Center
University of Michigan

During the past two decades, U.S. researchers, policymakers, and journalists have expressed concern that the nation's schools are failing to prepare students to meet the demands of the modern global economy. Researchers have interpreted international assessments as revealing serious weaknesses in mathematics and science proficiency (see, for example, Beaton et al.1996; Medrich and Griffith 1992; NCES 1995, 230–231). Although such claims can be strongly contested (c.d., Rotberg 1998), they support a broader climate of malaise, and even crisis, concerning the performance of U.S. schools.

In this climate, calls for reform and accountability at every level of the education system have taken on greater urgency. The stakes are often high: students in Chicago must pass a citywide test to be promoted to the next grade; students in Michigan can obtain endorsed diplomas only by passing the state's proficiency test; teachers with high-scoring classrooms can obtain cash rewards in some districts; and school principals are held accountable for school mean achievement.

For comparisons at the state level, the key source of data is the Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP), "the Nation's Report Card" (c.f., Mullis et al. l992). Administered every two years (though in different subject areas at each administration), TSA enables cross-sectional comparisons among participating states in several subject areas at several grades and allows estimation of trends in student mean proficiency over time. Participation has grown to include more than 40 states and U.S. territories. But what are we to make of such comparisons between states?

Most "users" of the TSA would like to view state proficiency means as reflecting the effectiveness of educational provision, policy, and practice within each state. If so, TSA would provide direct evidence of the quality of each state's educational system. Talking to those involved in reform, for example, I have found it common to view California's performance on TSA in certain subject areas as direct evidence of the failure of reform in that state. Yet even a cursory examination of TSA data reveals that state demographic composition, including poverty levels and ethnic composition, is strongly associated with state mean proficiency—and state trends in proficiency are undoubtedly associated with state trends in demography. Thus, critics claim that state means are surrogates of demography more than indicators of educational effects. This criticism has led to many calls for statistical adjustment of state means on the basis of student social and ethnic background. Indeed, it is possible to compare states within strata defined by ethnic background and parental education (as in Mullis et al. 1992), but such within-stratum comparisons control background differences only roughly and do not take into account the extent to which a school's demographic composition creates a context affecting student performance.

The National Assessment Governing Board, which provides policy direction to NAEP, has resolutely rejected the notion of reporting statistically adjusted state mean proficiency. Board members fear that adjustments for student background will lower expectations for school systems serving disadvantaged students. There are also sound statistical reasons to be skeptical about adjustments. Suppose, for example, that we use a regression analysis to compute state mean residuals, that is, discrepancies between the actual state means and the means expected on the basis of student composition. Such residuals have often been interpreted as indicators of the "value added" by the schooling system. Yet, if the regression model fails to include key aspects of educational policy and practice, the estimates of the association between student composition and outcomes will be biased. The bias would arise because the quality of educational provision and student composition would be positively correlated, with the most advantaged students tending to be found in the schools with the most favorable resources, policies, and practices. Failing, then, to control for the quality of educational provision will inflate estimates of the contribution of student demography. This inflation, in turn, will lead to biased "value added" indicators. The result is an over-adjustment for demography, such that systems serving the most advantaged students will tend to look less effective than they are. However, the magnitude of the over-adjustment is impossible to assess in the absence of data on the quality of school policy and practice (see Raudenbush and Willms [1995] for a thorough discussion of this problem in the context of school evaluation).

Interpretation of state proficiency means is thus terribly risky. We cannot equate unadjusted state mean proficiency with educational effectiveness as many reformers wish, yet adjusted means set up low expectations for states serving poor students and are statistically untrustworthy.

The problem of interpreting the results of the TSA frames the pair of investigations I shall discuss in this paper.[1] The debate over the meaning of state mean proficiency reflects a longstanding debate about the sources of inequality in academic achievement in the United States. If inequality in family background is the key to inequality in educational outcomes, then inequality in aggregate family background ought to be key to understanding differences in state achievement means. On the other hand, if inequality in school quality is key to understanding inequality in individual outcomes, then aggregate school quality ought to explain state variation. Fortunately, NAEP provides some reasonable data at the level of both the student and the school to test these propositions.

Our first investigation, then, tested models for student math proficiency within each of the participating states of TSA. This may be likened to a "meta-analysis" in which each state's data provide an independent study of the correlates of math proficiency. We examined student social, ethnic, and linguistic backgrounds, and home educational resources as predictors of student proficiency. Yet our models simultaneously included indicators of educational quality: course-taking opportunities, school climate, teacher qualifications, and cognitive stimulation in the classroom. Our findings, reasonably consistent across states, supported both the "home effects" and the "schooling effects" explanations: the hypothesized explanatory variables related to student outcomes as expected. This exercise may be criticized as merely recapitulating

[1] The research reported here was funded by the National Assessment of Educational Progress Data Reporting Program of the National Center for Education Statistics (NCES) under a grant to Michigan State University. The views expressed herein do not represent the position of NCES. This paper summarizes and discusses findings from two papers: "Synthesizing Results from the Trial State Assessment," to appear in the *Journal of Educational Statistics;* and "Inequality of Access to Educational Opportunity: A National Report Card for Eighth Grade Math," *Educational Evaluation and Policy Analysis,* 20(4), 253–268. Authors of both papers are Raudenbush, S.W., Fotiu, R.P., and Cheong, Y.F. I wish also to thank Marcy Wallace for administering the many tasks associated with the analysis and Zora M. Ziazi for her work on data analysis.

decades of educational research, and not even with the best available data.[2] Yet TSA does offer the opportunity to compare results across states, for it is the only data set that contains a large, representative sample of students in each of many states.

Perhaps more importantly, the analyses within states bears directly on controversies surrounding accountability at the state level. Our key finding was that, while states vary substantially in unadjusted proficiency means, once we control for NAEP indicators of student background and educational quality, nearly all of the state variation vanishes. This makes sense, in that state-level policies (e.g., regulations, incentives, and aid) can presumably affect student outcomes only by affecting specific educational resources and practices at a more local level, i.e., within schools and classrooms. If those local resources and practices were fully controlled in our models, there would be no direct role for state policy to affect student achievement.

Yet once we verify that state differences almost entirely reflect variation in measurable aspects of student background and school quality, our focus logically shifts to these "correlates of proficiency." In particular, state differences in correlates of proficiency *that can be manipulated by policy* become especially salient. This led to our second investigation: a study of state-to-state variation in the provision of key educational resources, in particular those resources found consistently related to student outcomes across states.

We were especially interested in equality of access to those resources as a function of student social and ethnic background. Our logic was as follows: having found what many prior studies have found, i.e., that socially disadvantaged and ethnic minority students are at high risk of poor performance, we are inclined to ask about the extent to which these students have access to key resources for learning.

Our results were again not surprising, but nonetheless disconcerting: socially disadvantaged students and ethnic minority students (particularly African American, Hispanic American, and Native American students) are significantly less likely than other students to have access to favorable course-taking oppor-

[2] The cross-sectional data of the TSA do not enable the degree of control for prior student achievement that is possible in a longitudinal study such as NELS. Moreover, NAEP indicators of educational policy and practice are not nearly as refined as are those in NELS.

tunities, school climates, qualified teachers, and cognitively stimulating classrooms. However, what is new and perhaps unique is a second finding based on TSA: the *degree* of social and ethnic inequality of access to resources varies substantially by state. This finding led us to propose a novel "report card" for states based not on mean outcomes, but rather on the extent to which the schools in a state provide key resources for learning. Moreover, our report card allows examination not only of state differences in overall access to these resources, but also state differences in the extent to which access is equitable as a function of social background and ethnicity.

These analyses, while fruitful in our view, also reveal important limitations in data provided by the TSA. These limitations are not so much on the outcome side, where most attention has focused on the construction of NAEP, but rather on the input side. Indicators of student background and especially of key educational resources are currently quite limited in the TSA. For example, student socioeconomic status is indicated by parental education in our analyses. Indicators of parental occupation, income, eligibility for free lunch, and census-based indicators of neighborhood demographic condition, housing, etc., are absent. Regarding school-level organization, NAEP includes indicators of disciplinary climate, but no indicators of staff cohesion, control, and expectations, or of academic press. Indicators of cognitive stimulation in the classroom are few and do not constitute a meaningful or reliable scale. Hence, we settled on a single indicator: emphasis on reasoning during math instruction.

Given the limitations of NAEP indicators of student background, school organization, and instruction, our finding that NAEP indicators can account for nearly all the variation between states was a pleasant surprise. A more refined set of indicators would, however, provide more useful information to those who wish to use TSA, not just to "take the temperature" of the states, but to identify specific targets and strategies for interventions aimed at reducing inequality and thereby improving overall levels of student proficiency.

In the following pages, I aim first to sketch briefly the longstanding debate over sources of educational inequality and its implications for accountability at the state level. Second, I describe the first phase of our investigation: the modeling of student proficiency within states as a function of student background and educational resources. Third, I report results of our second investigation, which focuses on student access to key educational resources in the participating states. A sub-theme in the description of each phase involves challenges of analysis and measurement that also have important implications for future summaries and uses of data from the TSA.

Home and School Differences As Sources of State Inequality in Mathematics Proficiency

The debate about how to interpret the results from the TSA mirrors the longstanding debate about home and school sources of inequality in student outcomes. Social and ethnic inequality in achievement constitutes a troublesome and enduring aspect of schooling in the U.S. Large achievement gaps between students of high and low socioeconomic status (SES) and between European American students, on the one hand, and African American and/or Hispanic students, on the other, have been verified in every major national study of secondary students, beginning with Coleman et al. (1966). Yet researchers have offered contrasting explanations for such inequality.

Home Environmental Inequality

From one standpoint, the school is an essentially neutral learning environment passively allowing sharp inequality in home circumstances to translate into similar inequalities in learning outcomes. Families have long been known to vary substantially in their capacities to provide educational environments that foster school readiness and reading literacy (Fraser 1959; Wolf 1968). Such differences are linked to social status indicators, including income, parental occupation, and parental education (Coleman et al. 1966; Peaker 1967). Parents of high social status are more likely than parents of low social status to have the resources and skills needed to support their children's academic learning.

If this explanation were completely sufficient to understand observed achievement gaps, variation in student achievement between schools would simply reflect the varied home environments of students attending those schools. Policy interventions aimed at increasing equity might focus primarily on early interventions such as Head Start and on providing support for the families of the most disadvantaged children. Interventions at the classroom or school levels, though perhaps laudable for increasing mean achievement, would hold less promise for reducing inequality.

School Environmental Inequality

From an entirely different standpoint, schools are a much more active force, subjecting essentially similar children to dramatically different learning experiences and thereby actively recreating in each new generation a wide intellectual inequality that conforms to the wide inequalities in earnings and occupational prestige. Clear expositions of this view appear in Ryan (1971), Bowles and Gintis (1976), and Kozol (1991). Tracking (Oakes 1985, 1990), differential teacher expectations (Rosenthal and Jacobson 1968; Rist 1970), and varied school ethos or climate (Rutter et al. 1979), course requirements (Lee and Bryk 1989), teacher subject matter and pedagogical knowledge (Finley 1984; Rosenbaum 1976), and level of cognitive stimulation in the classroom (Page 1990; Rowan, Raudenbush, and Cheong 1993) are aspects of the schooling system often viewed as fostering unequal opportunity and outcomes.

If inequality of schooling were the sole determinant of inequality of educational outcomes, inequality in school mean achievement would reflect school differences in policy and practice. Not surprisingly, those who have emphasized the school as a causal agent in creating educational inequality, while often endorsing compensatory educational policies, have called for sweeping structural reforms in the provision of schooling. These include the elimination of tracking, school finance reform that would equalize spending across rich and poor districts (Berne 1994), and a recasting of teacher preparation to foster more favorable expectations and more cognitively stimulating instruction for currently disadvantaged students. If the “school effects” explanation were correct, such reforms would reduce or eliminate differences between schools in achievement.

The debate reviewed above leaves school differences in student mean outcomes open to vastly different interpretations. One observer might view an elevated school mean as simply reflecting an advantaged school composition; another would attribute this success to excellent school governance, organization, policy, and instructional practice. Those who study school effects seek to measure key aspects of both student composition and school process to assess the relative contributions of each and to isolate those contributors to achievement that reformers can modify (Fuller 1987; Lee and Bryk 1993). Causal inference in such studies is always perilous because student composition and school process are inevitably correlated. Thus, if either student composition or school process is not measured well and is still included in the analysis, estimates of both will be biased.

Given the difficulty of conducting sound studies of school effects, it is not surprising that schemes designed to hold schools accountable for their mean achievement levels have encountered intense criticism (Willms 1992). School means that are not adjusted for student composition will typically convey an overly negative picture of school process in those schools with the most disadvantaged students. However, incorporating adjustments for composition typically leads to underestimates of the effectiveness of schools having favorable student composition (Raudenbush and Willms 1995).[3]

Implications of the Debate for Interpreting State Variation in Outcomes

All of the difficulties in interpreting school differences in mean outcomes are amplified when interest focuses on state mean differences. First, state means are simply aggregates of school means—the same means that have been found difficult to interpret in all but the most careful studies. Second, while all of the problems associated with interpreting either unadjusted or adjusted school means are present in adjusting state means, others are added. For example, the association between student composition and school processes will vary from state to state, as we show below, making the problem of finding meaningful adjustments for student composition even more perplexing. And differences in state means will at least partially reflect differences in state policy. Such policy differences may also be correlated with school composition and school process, creating extra uncertainty about the sources of state variation.

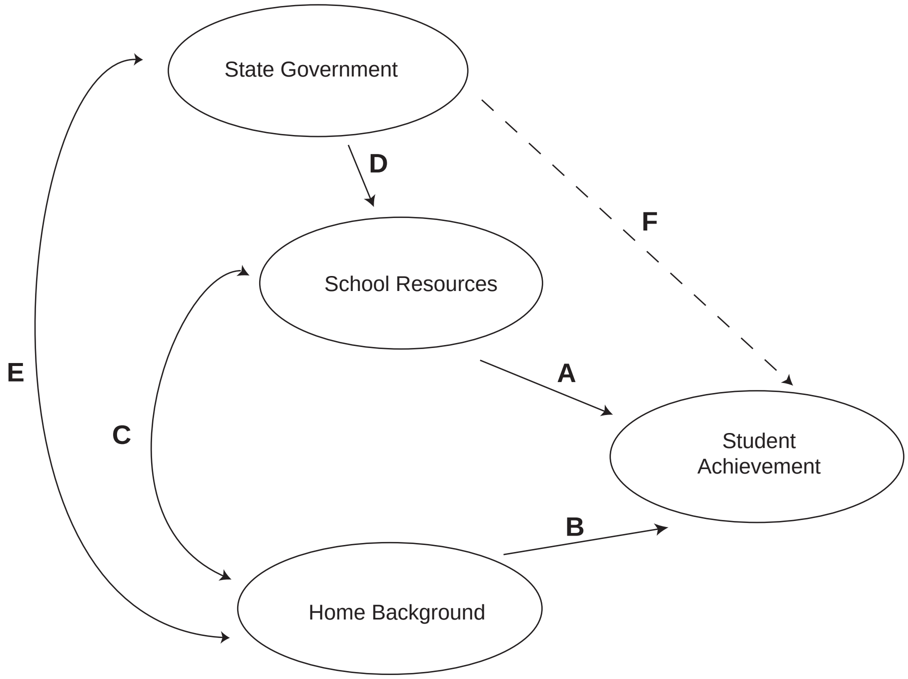
Thus, while making good estimates of state mean proficiency appears essential to any picture of the condition of the nation's education system, state differences in mean proficiency are, by themselves, intrinsically ambiguous at best and misleading at worst because of the inevitable temptation to make groundless causal inferences.

[3] Student advantage is typically positively correlated with effective school process. Analyses that control student demographics without incorporating good measures of school process will over-estimate the importance of student background, thus leading to overly severe adjustments for student background and thereby underestimating the effectiveness of schools serving advantaged students. Rarely do school accountability studies measure key aspects of school process.

The problem of interpreting state means can perhaps be clarified with reference to a simple causal model (figure 1). Those who interpret state means from TSA are typically interested in the role of state government in improving student achievement (arrow F of figure 1). However, in principle, states cannot directly alter student learning (which is why arrow F is a “dashed line” rather than a solid line). Instead, state policy may affect student achievement *indirectly* by encouraging favorable practice and resources at the level of the school or teacher (arrow D). Schools and teachers can directly affect student achievement (arrow A), though any analysis of such effects must account for student background (arrow B) because school and teacher practice are likely correlated with student background (arrow C).

The first phase of our analysis uses NAEP data to study arrows A and B, i.e., to assess contributing school and teacher quality and the contribution of student background in each of 41 states. The second phase considers arrow D,

Figure 1. Conceptual Model for State-level Policy Effect on Student Achievement



the differences between states with respect to those school and teacher resources and practices found consistently correlated with student achievement.

Phase I: Correlates of Proficiency within States

The first phase of our analysis was to study home and school correlates of eighth grade mathematics proficiency within each state. Our hypotheses were that student social, ethnic, and linguistic background, along with indicators of the home literacy environment, would be related to mathematics proficiency, as in past research; and that indicators of key aspects of school quality, such as course-taking opportunities, disciplinary climate, teacher qualifications, and cognitive stimulation in the classroom, would also predict proficiency. It was essential in this analysis that effects of student background and school quality indicators be adjusted for each other and for other contextual variables such as the composition of the school. This exercise could be viewed as much as a validation study of TSA indicators as a test of theory. We wanted to see whether TSA indicators of home background and school quality were sufficiently well measured to reproduce essential findings of past research. We also sought to examine the power of our within-state models to account for variation between states.

Our expectation was that key variables measured at the student and school level would account for most of the variation between states. This expectation was driven by substantive, rather than statistical, concerns. Controlling for explanatory variables at lower levels of aggregation, such as the student or the school, need not reduce variation at a higher level, such as the state. The adjusted between-state variation can, in principle, be either smaller or larger than the unadjusted between-state variation. However, it stands to reason that states will vary in outcomes for two reasons: selection processes and effects of state educational policy and practice. Selection processes arise because patterns of settlement, fertility, and economic dislocation produce state variation in the demographic and cultural backgrounds of students and their families. Educational policies and practices of schools vary because of the uniquely decentralized character of the U.S. education system and because states and localities tailor the provision of education to the populations they serve. However, states are limited in the “levers” available to them to affect student outcomes. These levers include regulations, incentives, and forms of aid that can have only indirect effects on students by affecting district and school leadership and, ultimately, instruction. It follows that if key aspects of selection, school practice, and instruction are controlled, no state variation will remain to be explained. In terms of figure 1, once arrows A and B are controlled, arrow F should be nonsignificant. This makes sense theoretically but may be difficult to show empirically with NAEP data because NAEP indicators of school resources and home background are limited.

Sample and Measures

Sample

The analyses are based on data from 99,980 eighth graders attending 3,537 schools located in the 41 states and territories participating in the 1992 Trial State Assessment in mathematics. Thus, the average state sample included 2,377 students and 86 schools.

Students within each state were selected by means of a two-stage cluster sample with stratification at the first stage. Specifically, schools were first stratified on the basis of urbanicity, minority concentration, size, and area income; then (a) schools were selected at random within strata with a probability proportional to student grade level enrollment; and (b) students were systematically selected from a list of students, given a random starting point, within schools. It is essential that the analysis plan take into account the stratified and clustered nature of the sample.

Measures

Table 1 lists the variables used and their descriptive statistics. The variables include student outcome data, demographic indicators, home environmental indicators, and classroom and school characteristics.

Measures of math proficiency. The math proficiency data collected as part of NAEP involve a matrix-sampling scheme in which each student was observed on only a subset of relevant items. Rather than yielding a single measured variable, NAEP produces five “plausible values”—random draws from the estimated posterior distribution of each student’s “true” outcome given the subset of items and other data observed on that student (Johnson, Mazzeo, and Kline 1993).

Measures of student demographics. Student demographic variables consist of gender (indicator for male), ethnicity (indicators for Hispanic American, non-Hispanic black American, Asian American, and Native American, with

Table 1. Descriptive Statistics for Student- and School-level Variables for the Combined Sample

Variables	Code and range	Mean	Standard deviation
Student-level data (99,980 students)			
Outcome variables			
Math proficiency 1	(-2.96, 3.06)	0.03	0.99
Math proficiency 2	(-3.82, 2.71)	0.03	0.99
Math proficiency 3	(-3.75, 3.33)	0.03	0.99
Math proficiency 4	(-3.22, 2.87)	0.03	0.99
Math proficiency 5	(-3.84, 2.76)	0.03	0.99
Demographics			
Male	0 = No, 1 = Yes	0.50	0.51
African American	0 = No, 1 = Yes	0.15	0.36
Hispanic American	0 = No, 1 = Yes	0.14	0.35
Asian American	0 = No, 1 = Yes	0.03	0.19
Native American	0 = No, 1 = Yes	0.02	0.12
Not born in U.S.	0 = No, 1 = Yes	0.07	0.26
Student-level data (99,980 students)			
Home environment			
Living with both parents	0 = No, 1 = Yes	0.70	0.47
Living with one parent	0 = No, 1 = Yes	0.20	0.41
Parental education—high school diploma	0 = No, 1 = Yes	0.30	0.47
Parental education—more than high school diploma	0 = No, 1 = Yes	0.18	0.40
Parental education—bachelor’s degree or more	0 = No, 1 = Yes	0.26	0.45
Hours watching TV	(0, 6)	3.17	1.61
Changed school in past 2 years	0 = No, 1 = Yes	0.22	0.42
Get newspaper regularly	0 = No, 1 = Yes	0.73	0.46
More than 25 books in home	0 = No, 1 = Yes	0.91	0.29
Get magazines regularly	0 = No, 1 = Yes	0.76	0.44
Classroom characteristics			
Taking algebra	0 = No, 1 = Yes	0.19	0.40
Taking pre-algebra	0 = No, 1 = Yes	0.25	0.44
Teaching experience of math teacher	(1, 30)	13.44	8.85
Math teacher majored in math	0 = No, 1 = Yes	0.43	0.51

Table 1. Descriptive Statistics for Student- and School-level Variables for the Combined Sample (continued)

Variables	Code and range	Mean	Standard deviation
Math teacher majored in math education	0 = No, 1 = Yes	0.18	0.39
Math teacher did graduate work	0 = No, 1 = Yes	0.47	0.51
Math teacher emphasized reasoning/ analysis in class	0 = otherwise 1 = heavy/moderate	0.46	0.51
School-level data (3,537 schools)			
School-level variables			
Median income (in thousands)	(9.073, 85.567)	28.80	10.73
Instructional dollars per pupil	(7.5, 17.5)	67.22	30.23
Percent minority	(1,100)	28.02	27.70
Urban location	0 = No, 1 = Yes	0.23	0.42
Rural location	0 = No, 1 = Yes	0.23	0.42
Offering 8th grade algebra for high school credits	0 = No, 1 = Yes	0.75	0.43
Availability of computer	0 = No, 1 = Yes	0.83	0.37
School climate	(-3.003, 1.191)	0.00	0.63

European American as the reference group), national origin (indicator for born outside the U.S.), family type (indicators for living at home with a single parent, living at home with both parents, with other type as the reference group), and parental education (indicators for high school graduate, some education after high school, and college graduate, with not graduated from high school or the eighth grader not knowing parents' educational level as the reference group).

Table 1 presents the descriptive statistics on student demographics for the combined 41 states. As table 1 shows, half of the 99,980 students were male. African Americans made up 15 percent of the sample; Hispanic Americans, 14 percent; Asians, 3 percent; and Native Americans, 2 percent; and 7 percent of the students were not born in the U.S. In addition, 70 percent of the students indicated that they had two parents residing at home, and 20 percent of students reported that they lived in a single-parent household. For 30 percent of the sample, either the mom or the dad held a high school diploma; for 18 percent, one parent had some education after high school graduation; and for 26 percent, at least one parent graduated from college.

Measures of home environment. Home environment variables include amount of time watching television, mobility (as indexed by whether a student changed schools in the past two years), home literacy environment (indicators for receiving a newspaper, having more than 25 books, and subscription of magazines). Table 1 indicates that the students spent 3.17 hours daily on average watching TV. Less than a quarter of them (22 percent) reported that they had changed schools in the past two years. About three-fourths of the students (73 percent and 76 percent) indicated that their households regularly got newspaper and magazines, respectively. The great majority of the students, 91 percent, had more than 25 books in their homes.

Measures of classroom characteristics. Classroom characteristics involve type of course (indicators for pre-algebra, algebra, with other course as the reference group), the teaching experience and qualifications of the teacher of the student (indicators for undergraduate math major in college, math education major in college, with other major as the reference group; and an indicator for having a graduate degree), as well as teacher-reported emphasis on reasoning in the classroom (an indicator for moderate to high emphasis). The data on teacher background and pedagogical practice were taken from responses to questionnaires administered to the mathematics teachers of the students sampled.

Table 1 shows that 19 percent of the students in the sample enrolled in an algebra course and 25 percent of them took pre-algebra. The average number of years of teaching experience for the teachers of the students sampled was about 13. Furthermore, 43 percent of the students had a teacher who majored in mathematics as an undergraduate; 18 percent of the students had a teacher who was a math education major; and 47 percent of the students had a teacher who got a graduate degree. About half of the students (47 percent) attended a classroom where reasoning received moderate to high level of emphasis.

Measures of school characteristics. School characteristics include the social and racial composition of a school as measured by median income and percent minority (Hispanic and African American students). Other school-level measures are location (indicators for an urban school, a rural school, with suburban school as a reference group), and financial and computing resources as indexed by instructional dollars per pupil and availability of computers (an indicator for the availability of computers in a math classroom or a lab for most of the time), course offerings (an indicator for the availability of algebra for high school credit), and a scale measuring the disciplinary climate of the school. The scale was created from the following items indicating the extent to which each was a problem in the school: tardiness, absenteeism, cutting classes, physical conflicts, drug and alcohol use, health, teacher absenteeism, racial or cultural conflict. Each item was first standardized, and the scale was constructed as the average of the nine standardized scores. Average Cronbach's alpha for the 41 states was .79.

Analytic Approach

Math Proficiency

Our strategy for modeling math proficiency has two stages: a within-state analysis and a between-state analysis. The within-state analysis uses a hierarchical linear model to handle the clustered character of the sample. Sample design weights are applied at the student level to accommodate the stratified character of the sample and the associated over-sampling of certain subgroups. This analysis is replicated for each plausible value and the results pooled as recommended in Little and Schenker (1994) and Mislevy (1992), using a specialized version of the HLM program (Bryk, Raudenbush, and Congdon 1994) originally adapted for multiple plausible values by Arnold, Kaufman, and Sedlacek (1992). The output for each state is a vector of parameter estimates and their estimated sampling variance matrix. These then provide input data for the second stage of the analysis, which involves an empirical Bayes and a Bayesian synthesis of findings across states. The syntheses employ the method of moments (Raudenbush 1994) and the Gibbs sampling (Gelfand and Smith 1990). (See Raudenbush, Fotiu, and Cheong [1998] for a full exposition of the approach.) Taken together, the two stages have the structure of a planned "meta-analysis" (Glass 1976) in which each state's separate analysis constitutes a "study," and the between-state analysis combines these results.

Within-state Models

To address these questions, we first formulated within each state two separate two-level hierarchical models, one with and one without covariates (measures on student demographics, home environment, and classroom and school characteristics). Past research on the associations between the social distribution of educational resources and outcomes guided the specification of the former model (e.g., Bernstein 1970; Bryk and Thum 1989; Coleman et al.1966; Finley 1984; Oakes 1985; Page 1990; Raudenbush, Rowan, and Cheong 1993; Rosenbaum 1976; and Rutter et al. 1979). The model is

$$1. \qquad Y_{ijk} = \beta_{0k} + \sum_{p=1}^{p} \beta_{pk} X_{pijk} + u_{jk} + e_{ijk} \,,$$

where

Y_{ijk} is the math proficiency score for student *i* in school *j* and state *k*;
β_{0k} is the mean for state *k*, which is adjusted for the school- and student-level covariates;
X_{pijk} is the p^{th} covariate, which is centered around the Michigan mean;
β_{pk} is the regression coefficient associated with each X_{pijk};
u_{jk} and e_{ijk} are the residual random school and student effects. They are assumed independently and normally distributed with ω_k^2 and σ_k^2 respectively.

Estimates of the two variance components, ω_k^2 and σ_k^2, incorporate variation associated with the cluster sample so that the maximum likelihood (ML) estimate of each regression coefficient and its standard error incorporates the extra variation arising from the clustered nature of the sample. The use of sampling weights accounts for unequal probability of selection and multiple plausible value analysis accounts for the estimation of proficiency.

Deviating the school- and student-level covariates around the Michigan means allows us to obtain more precise estimates of various parameters for our own state, Michigan.[4] For the sake of simplicity, we forego the option of allowing any of the partial effects associated with student-level covariates to vary randomly from school to school within state *k*. Thus, only β_{0k}, the intercept, varies randomly across schools within states.

Between-state Models

The between-state synthesis combined the output produced by each state to obtain inferences on parameters for individual states as well as global parameters. The output from the within-state analysis for state *k* consisted of the ML estimates b_k of the state mean and its estimated sampling variance v_k. The estimate b_k is assumed to vary around its corresponding parameter β_k with an

[4] The covariates can be deviated around other constants such as the national means for other purposes.

unique error r_k associated with the sample for state *k*, which has a known sampling variance v_k, i.e.,

2. $b_k = \beta_k + r_k , r_k - N(0,v_k)$.

The parameter β_k is in turn assumed to vary around an overall mean γ plus a random error associated with state *k*, μ_k. We may write

3. $\beta_k = \gamma + \mu_k , \mu_k - N(0,\tau_k)$.

The random error has a variance of τ.

Table 2 lists the approximate posterior means and standard deviations of the various regression coefficients, and the estimates of between-state variance and their square roots.[5] We computed z-ratios for the regression coefficients to evaluate the null hypothesis that a particular regression coefficient pooled across states was 0. A z-ratio larger than 2 or 3, as indicated by asterisks in table 2, lent support to rejection of the null hypothesis.

Student demographics. Controlling for home environments and for classroom and school characteristics, the results suggest that, on average, males had higher scores than females; and African Americans, Hispanic Americans, and Native Americans exhibited lower proficiency than did European or Asian Americans. For instance, African Americans obtained, on average, about half a standard deviation lower math proficiency than did European Americans. Net of other covariates, students who were born in the United States scored higher that those who were not. The partial effects associated with the African American and Hispanic American ethnicity and the place of birth variables seem to vary from state to state.

Home environment. Controlling all other covariates, family structure, parental education, and home literacy environment were related to proficiency. Students who lived with either one parent or both parents outperformed those who did not and also those who did not know the educational levels of their parents. Students whose parents had education beyond high school and those whose parents had college degrees scored higher than did those whose parents had not graduated from high school. Furthermore, students coming from house-

[5] Table 2 gives the empirical Bayes summary results. Raudenbush et al. (in press) provide results from the fully Bayesian synthesis and compared the two sets of results. Individual state results are available upon request.

Table 2. Empirical Bayes Summary of State-by-State Results

Predictors	Approximate posterior mean of γ_p	Approximate posterior standard deviation of γ_p	Estimate of between-state variance, τ_p	Square root of the estimate of between-state variance, $\tau_p^{1/2}$
Demographics				
Male	0.0904*	0.0061	0.0005	0.0231
African American	-0.4583*	0.0215	0.0123	0.1107
Hispanic American	-0.3894*	0.0271	0.0239	0.1546
Asian American	0.1288*	0.0216	0.0032	0.0565
Native American	-0.2162*	0.0223	0.0043	0.0654
Not born in U.S.	-0.2369*	0.0211	0.0110	0.1046
Home environment				
Living with both parents	0.2884*	0.0116	0.0021	0.0456
Living with one parent	0.2500*	0.0124	0.0022	0.0471
Parental education—high school diploma	0.0567*	0.0082	0.0008	0.0273
Parental education—more than high school diploma	0.2455*	0.0085	0.0217	0.2455
Parental education—college degree	0.2146*	0.0125	0.0041	0.0638
Hours watching TV	-0.0404*	0.0024	0.0001	0.0112
Changed school in past 2 years	-0.0640*	0.0063	0.0000	0.0000
Get newspaper regularly	0.0277*	0.0057	0.0000	0.0000
More than 25 books in home	0.2051*	0.0092	0.0000	0.0000
Get magazines regularly	0.1006*	0.0075	0.0007	0.0259
Classroom characteristics				
Taking algebra	0.9830*	0.0201	0.0141	0.1188
Taking pre-algebra	0.3972*	0.0159	0.0083	0.0912
Teaching experience of math teacher	0.0029*	0.0006	0.0000	0.0000
Math teacher majored in math	0.0844*	0.0121	0.0038	0.0844

Table 2. Empirical Bayes Summary of State-by-State Results (continued)

Predictors	Approximate posterior mean of γ_p	Approximate posterior standard deviation of γ_p	Estimate of between-state variance, τ_p	Square root of the estimate of between-state variance, $\tau_p^{1/2}$
Math teacher majored in math education	0.0823*	0.0149	0.0055	0.0738
Math teacher did graduate work	0.0101	0.0084	0.0010	0.0320
Math teacher emphasized reasoning/analysis in class	0.1373*	0.0096	0.0023	0.0478
School characteristics				
Median income	0.0059*	0.0007	0.0000	0.0000
Instructional dollars per pupil	0.0000	0.0000	0.0000	0.0000
Percent minority	-0.0036*	0.0000	0.0000	0.0000
Urban location	0.0140	0.0143	0.0014	0.0380
Rural location	-0.0191	0.0225	0.0125	0.1120
Offering 8th grade algebra for high school credits	-0.0425*	0.0138	0.0018	0.0428
Availability of computer	0.0024	0.0124	0.0000	0.0000
School climate	0.0378*	0.0079	0.0000	0.0000
Intercept				
Intercept	0.0680	0.0096	0.0000	0.0000

*** z-score > 3.**

holds that had more than 25 books in the home and received newspaper and magazines regularly had higher math proficiency than those who came from households that did not. There were statistically significant negative partial effects associated with time spent watching TV and changing school in the past two years. Three of the between-state variance estimates were 0.

Classroom characteristics. Enrollment in algebra and pre-algebra were positively related to math scores, all else being equal. Those who took algebra scored about one standard deviation higher than the reference group, whose students took eighth grade math or other non-algebra course or who did not take any math course. Those who enrolled in pre-algebra scored about 0.4 standard deviation higher than the reference group. Teaching experience, teacher subject matter expertise (as indicated, respectively, by majoring in math or math education), and emphasis on reasoning[6] were also positively correlated with proficiency in math, net of the effects of other covariates.

School characteristics. School composition effects were manifest, net all other predictors, including student demographic background. In particular, school median income was positively related to proficiency, and percent minority was negatively related to proficiency. Thus, school social class and ethnic segregation effects tend to reinforce differences based on individual social class and ethnicity. All else being equal, a favorable school climate was positively related to proficiency. The estimated partial effect of school algebra was statistically significant and negative. Note that this effect represented the expected difference in math proficiency between a student not taking algebra in a school that offered algebra and a student in a school that did not offer algebra. One implication of the predominantly negative effect across the states is that there are at least some students in schools not offering algebra who would have benefited from enrollment in an algebra course had they attended schools that did offer algebra. In addition, as taking algebra was, in general, the most powerful single predictor of proficiency, one must conclude that attending a school that offers algebra is related positively to math proficiency.

In sum, the relevant covariates include indicators of student demographic status, home environment, and school composition; these relate to proficiency as expected. At the school level, a curriculum that includes opportunities to take high school algebra and a positive climate were linked to proficiency. At the classroom level, teachers' subject-matter preparation, as indicated by hav-

[6] One would expect the level of reasoning to increase with teacher's education (e.g., a teacher's undergraduate major) and the difficulty of the course (e.g., an algebra course versus a general mathematics course). Emphasis on reasoning, teacher's education, and course type thus may jointly influence math proficiency. To understand how these various predictors may be correlated with the math scores, two models were specified, one with and one without emphasis on reasoning entered as a predictor. The results showed that reasoning, independent of all other covariates, was positively related to math proficiency. In fact, the estimates of other predictors remained nearly the same in the two models.

ing majored in math or in math education, and emphasis on mathematical reasoning predicted elevated proficiency.

Variance Reduction

Figures 2 and 3 give the approximate marginal posterior for the variance for the intercept, that is, for var(β_{0k}) = τ for the unconditional (with no covariates) and conditional models.[7] Figure 3 shows unmistakable evidence of heterogeneity between states (note that 0 is not a plausible value for τ). However, there is considerable uncertainty about the magnitude of this heterogeneity.

The math proficiency measure was on a scale with a mean near 0 and a variance of approximately unity. The posterior mean of τ is .088, implying that about 8.8 percent of the variance in the outcome lies between states. However, τ values as small as .04 and as large as .14 are not improbable. Thus, it appears that from 4 percent to 14 percent of the variance in the outcome lies between states.

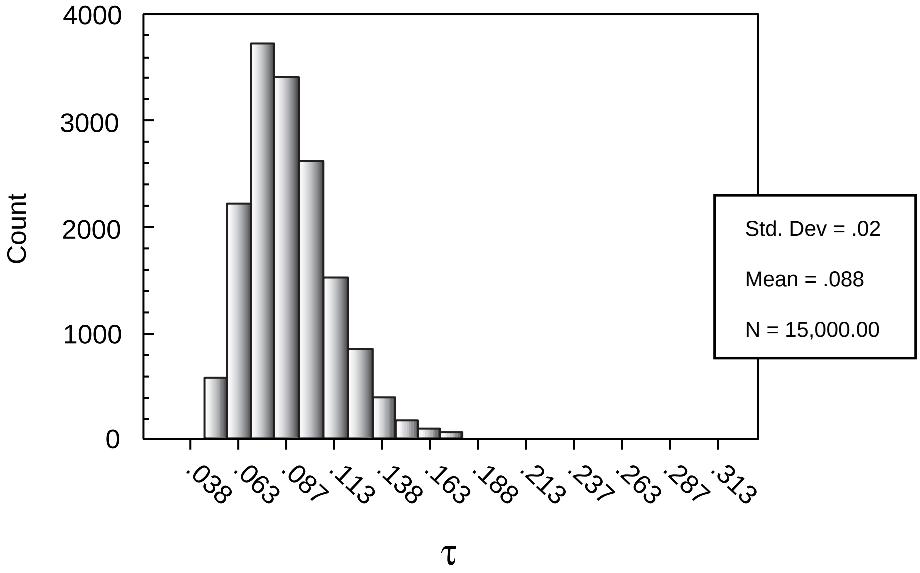
Whereas figure 3 shows evidence of heterogeneity between states (note that 0 is not a plausible value for τ) after controlling for the various measures, there is every reason to believe that the magnitude of this heterogeneity is small. The posterior mean of τ is .018, implying that 1.8 percent of the variance in the outcome lies between the intercepts of the states. Moreover, the unknown value of τ is unlikely to exceed .03 or 3 percent of the total variance in the outcome. It appears that from .004 percent to 3 percent of the variance in the intercept lies between states after controlling for covariates. Thus, most of the state-to-state heterogeneity is explainable on the basis of covariates defined on students, teachers, and schools. This indicates, in general, that states with high mean proficiency tend to be advantaged on the relevant covariates and that these advantages account for most state-to-state variation in proficiency.

Phase II: Inequality of Access to Educational Opportunity

In terms of figure 1, our "first phase" analysis found certain school resources (arrow A) and student background indicators (arrow B) to be quite

[7] The figures are output obtained from the Bayesian synthesis (see Raudenbush, Fotiu, and Cheong [in press] for a description of the approach).

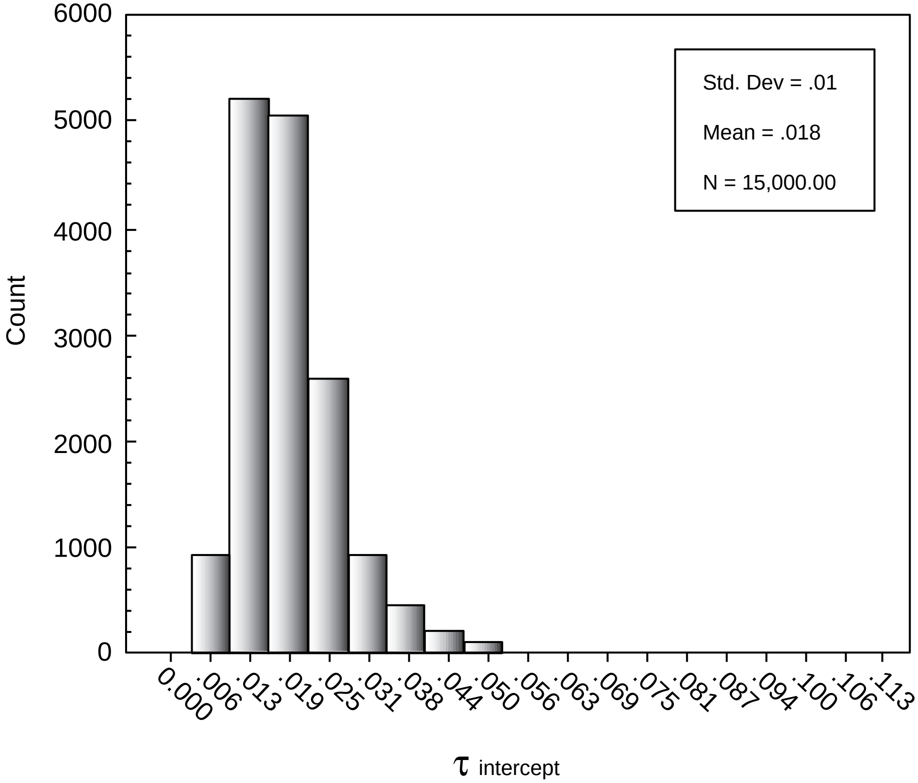
Figure 2. Estimated Posterior Distribution of τ: Unconditional Model



consistently related to student achievement. Controlling for these, state differences in achievement (arrow F) became small, perhaps negligible. This encouraged us to abandon further investigation of state means, whether adjusted or unadjusted. Rather, we sought in Phase II of our investigation to examine state differences in school resources. Given the consistent association between advantaged home background and achievement, we were especially interested in the equity with which the school resources are distributed. We asked: "Does the distribution of school resources likely reinforce or counteract inequalities arising from home environment? Do states differ, not only in the provision of resources, but also in the equity with which they are distributed?"

One product of this work is a different kind of "report card" for states than is typically made available to policymakers. The typical report card provides unadjusted differences between states in academic proficiency. This typical report card, though conveying some useful information, can easily mislead. It tends to provide an overly negative portrayal of education systems in states with comparatively disadvantaged demographics and an overly rosy picture of

Figure 3. Estimated Posterior Distribution of τ intercept: Conditional Model



education in states with more advantaged students. Moreover, it provides little insight into ways in which policy changes might produce better outcomes.

The report card we present compares states on educational opportunities, resources, or processes theoretically and empirically linked to outcomes. It reveals the equity with which these are distributed as a function of student social background and ethnicity. It therefore points the discussion toward interventions that would increase the quality and equity of education provision.

In modeling the relationship between student demographic background and educational resources, our analysis strategy depended on whether the educational resource in question was measured dichotomously or continuously. Dichotomous resources included school course offering (1 = school offers high school algebra, 0 = school does not offer high school algebra) teacher education (1 = teacher majored in math, 0 = teacher did not major in math), and emphasis on reasoning in the classroom (1 = high, 0 = other).

Model for the Continuous Outcome (Disciplinary Climate)

The method of estimation for the model studying school climate involves a two-level hierarchical linear model (Bryk and Raudenbush 1992) with students nested within states. Robust standard errors were computed using the generalized estimating equation approach of Zeger, Liang, and Albert (1988). These standard errors are relatively insensitive to mis-specification of the variances and covariances at the two levels and to the distributional assumptions at each level. State-specific effects were estimated via empirical Bayes (Morris 1983; Raudenbush 1988).

Specifically, we estimated a within-state model in which ethnicity, parental education, and the ethnicity-by-parent interaction predicted school climate. Ethnicity was represented by four dummy variables and parental education by two dummy variables. Allowing for the ethnicity-by-parent interaction effect enabled us to model access to resources for each sub-group (e.g., African Americans of low, middle, or high parental education). We allowed coefficients for the parental education dummies and for African American and Hispanic American ethnicity to vary randomly over states, thus allowing state-by-state comparisons. Sample sizes of Asian Americans and Native Americans were, unfortunately, too small to allow such a fine-grained analysis.

Models for the Dichotomous Resource Indicators

The same explanatory model for the school climate was specified for each dichotomous outcome. In this case, however, we used a two-level logistic regression model, estimated by penalized quasi-likelihood (Breslow and Clayton 1993), with robust standard errors. Such a model is equivalent to a 2 by 3 by 5 by 41 contingency table with 2 levels of the outcome, 3 levels of parent education, 5 levels of ethnicity, and 41 levels representing states.

Results

We now consider the degree of ethnic and social equality in access to the four resources of interest. Specifically, we ask the following questions for each resource indicator:

1. Averaging within the 41 states participating in the TSA, to what extent does student social background, as indicated by parental education and student ethnicity, predict access to the resources?
2. Does the degree of inequality in access vary by state? If so, how do the 41 states compare?

Results Averaged Across States

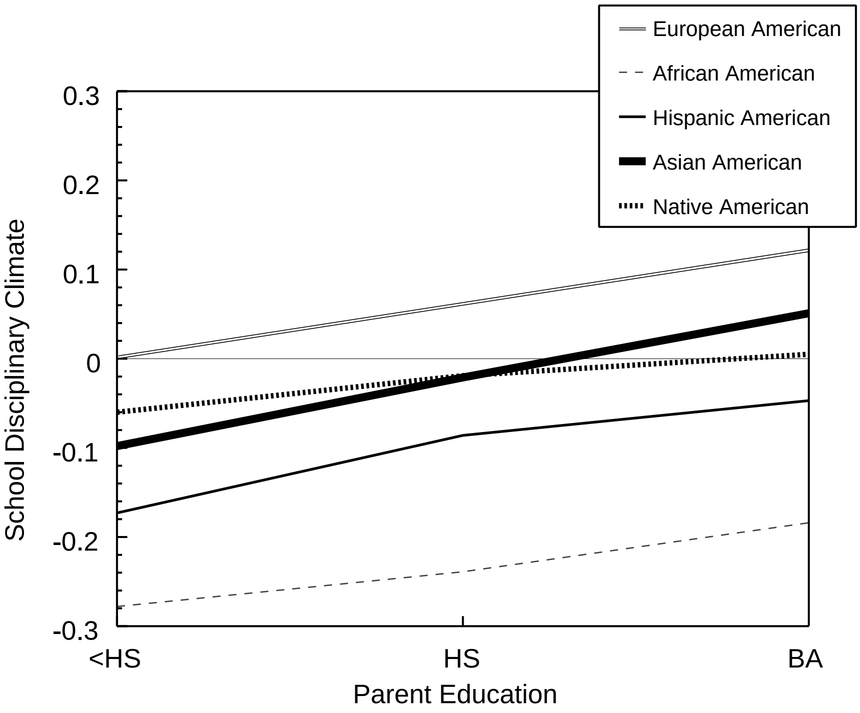
School Disciplinary Climate

Figure 4 gives the graph of the fitted model in which ethnicity and parental education predict access to favorable disciplinary climate. The figure shows that higher levels of parental education are clearly linked to more favorable disciplinary climate. The near parallelism of the five lines (with the exception of the line for Native Americans, which is based on a comparatively small sample) reflects the absence of any statistical evidence of a two-way interaction involving parental education and ethnicity. There is a substantial significant vertical displacement between ethnic groups. Pairwise comparisons using a Bonferroni adjustment to control the family-wise Type I error rate at the 5 percent level indicated four separate clusters of means (in descending order of magnitude): (a) European Americans; (b) Asian Americans and Native Americans; (c) Hispanic Americans; and (d) African Americans. Given that the school climate outcome had a mean of 0 and a standard deviation of 0.63, the differences manifest in figure 4 are non-trivial in magnitude: About 0.20 standard deviation units separate those with parents having a BA from those whose parents were without a high school diploma; nearly half a standard deviation separates European Americans and African Americans.

Access to High School Algebra

Figure 5 plots the predicted probability of attending a school that offers high school algebra for eighth graders as a function of parental education for

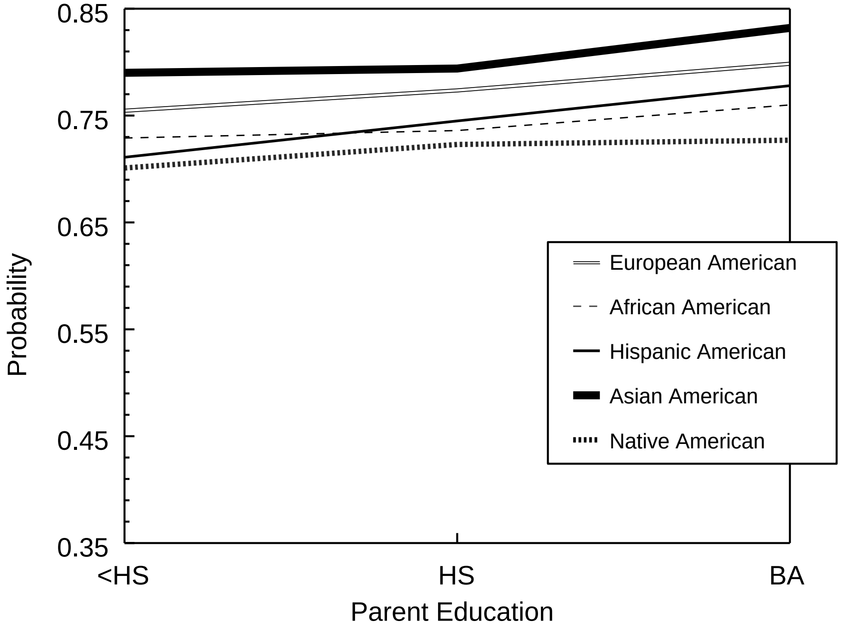
Figure 4. Predicted School Disciplinary Climate as a Function of Parent Education and Ethnicity



each of the five major ethnic groups under study. We see that parental education is positively associated with the probability of attending such a school. As in the case of climate, the near parallelism of the five lines reflects the absence of any statistical evidence of a two-way interaction involving parental education and ethnicity. Again, we find a significant vertical displacement between ethnic groups. Pairwise comparisons using a Bonferroni adjustment to control the family-wise Type I error rate at the 5 percent level indicated three separate clusters of ethnic group probabilities (in descending order of magnitude): (a) Asian Americans; (b) European Americans, African Americans, and Hispanic Americans; and (c) Native Americans. The differences manifest in figure 2 are comparatively modest in magnitude.

The regression coefficients for the predictors give the associated partial effects in terms of log-odds. Besides computing predicted probabilities based on the regression coefficients, one could compute odds ratios as well. For in-

Figure 5. Predicted Probability of Assignment to a School That Offers Algebra as a Function of Parent Education and Ethnicity



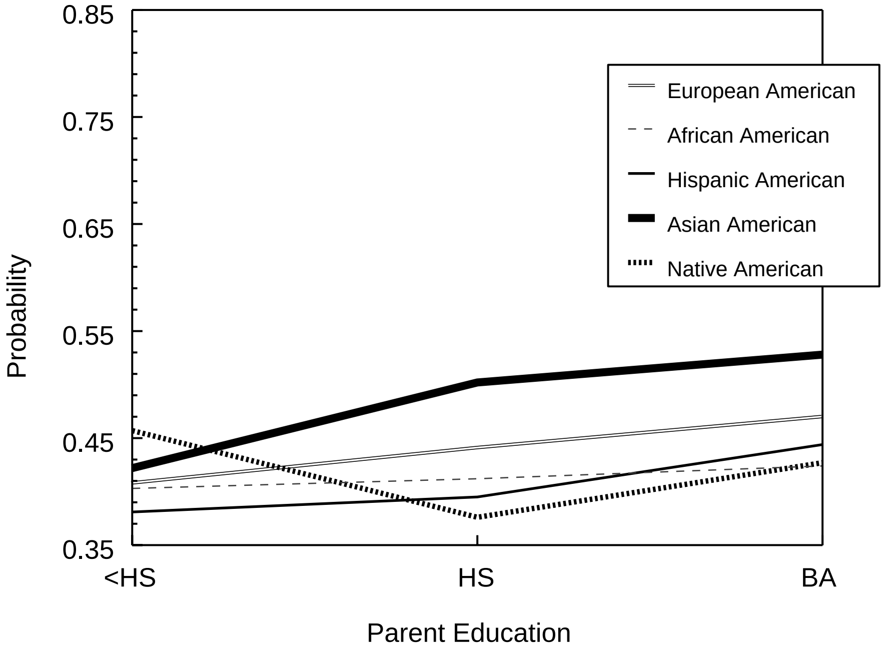
stance, the odds ratio of offering algebra for a school attended by a student whose parent had college education versus a school attended by a student whose parent had less than high school education is exp$\{\acute{\alpha}_{BA}\}$ = exp{-0.244} = 0.784.

We now turn to two classroom-level resources for learning: teacher subject matter preparation, as indicated by having majored in mathematics, and a cognitively stimulating environment, as indicated by an instructional emphasis on mathematical reasoning. In both cases, we find that social background (as indicated by parental education) and ethnicity are linked to access to the resource. However, the findings are more complex than those reported above, in that a two-way interaction is manifest in the case of these two classroom-level resources.

Teacher Preparation

Figure 6 plots the predicted probability of encountering a math teacher who majored in math as a function of social background and ethnicity. The figure shows that higher levels of parental education are linked to a higher

Figure 6. Predicted Probability of Assignment to a Teacher Who Majored in Math As a Function of Parent Education and Ethnicity

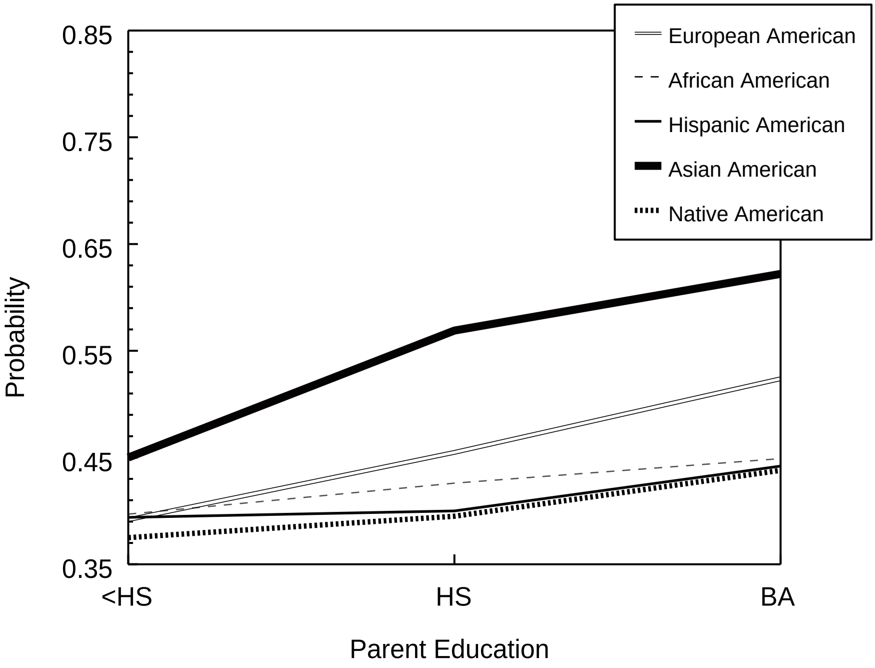


probability of encountering such a teacher. However, the magnitude of this relationship depends upon ethnicity. The link between social background and teacher preparation is strongest for Asian Americans and European Americans and weakest for African Americans, Hispanic Americans, and Native Americans. Equivalently, we can say that ethnic gaps in access to the resource are manifest, but are more pronounced at higher than at lower levels of parent education.

Emphasis on Reasoning

Figure 7 plots the predicted probability of encountering a math teacher who emphasizes mathematical reasoning during instruction. Again there is a positive relationship between parent education and this probability, but again the magnitude of this association depends upon ethnicity. The link between parental education and access to reasoning is strongest for Asian Americans and European Americans and weakest for the other three groups. Equivalently,

Figure 7. Predicted Probability of Assignment to a Math Teacher Who Emphasizes Reasoning As a Function of Parent Education and Ethnicity



just as in the case of teacher preparation, we can say that ethnic gaps in access to the resource are manifest, but are more pronounced at higher than at lower levels of parent education.

Summary

In sum, we find evidence of ethnic and social inequality in access to all four resource indicators when averaging across the 41 states. Main effects of both ethnicity and social background generally parallel previous findings in predicting student achievement. Thus, just as high parental education predicts favorable outcomes, it also predicts access to schools with favorable climates, schools that offer algebra, teachers with training in mathematics, and classrooms that emphasize reasoning. Similarly, ethnic groups disadvantaged in outcomes (African Americans, Hispanic Americans, and Native Americans) also encounter less access to these resources for learning.

State Variation in Access to Resources

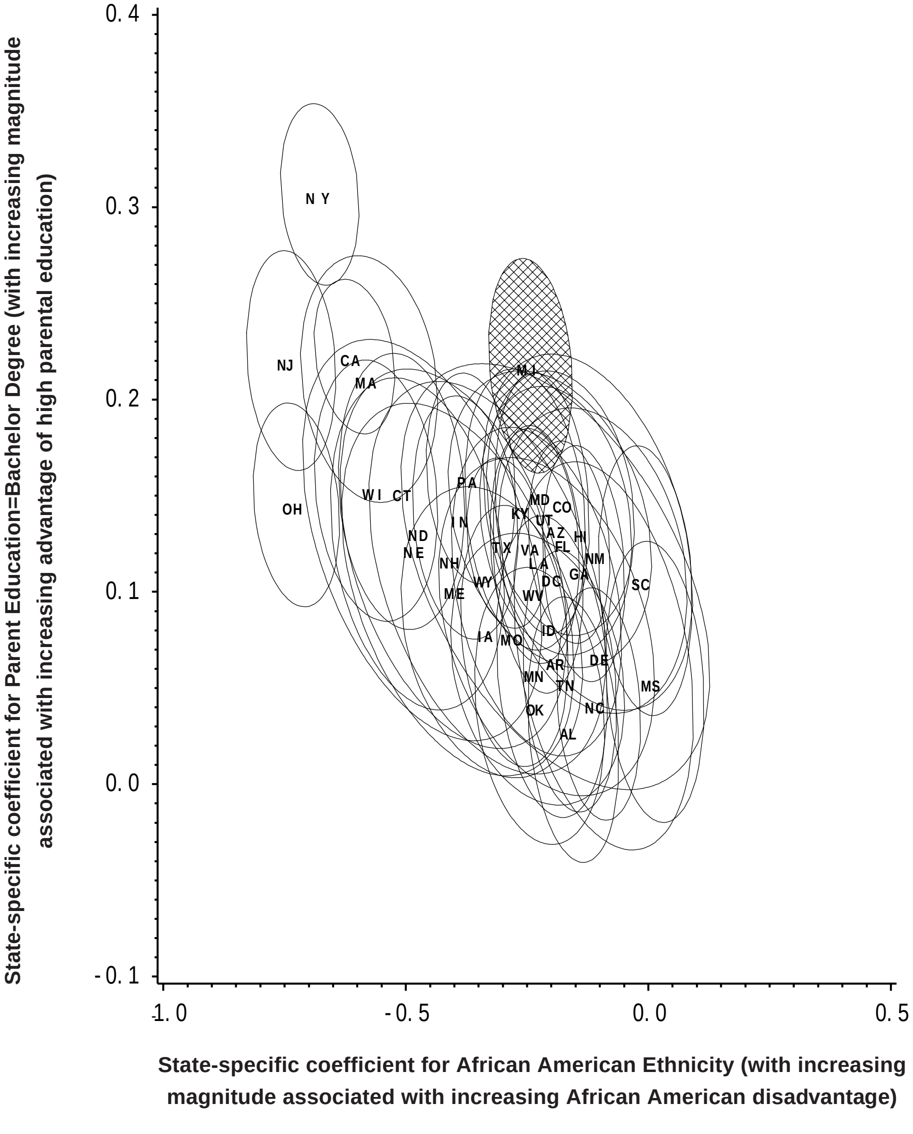
The pooled, within-state findings regarding social and ethnic inequality in access to a favorable school climate provide an “on-average” picture of inequality in access to resources over 41 states. However, these on-average results poorly represent the picture that we find in many states. In fact, the data reveal substantial evidence of state variation.

The case of school disciplinary climate illustrates the substantial variation across states. Figure 8 plots 95 percent bivariate confidence ellipses for the 41 states where the vertical axis is social inequality (as indicated by mean gaps in school climate between students having parental education of BA and less than high school) and the horizontal axis is ethnic inequality (as indicated by mean differences between African Americans and European Americans.[8] Four features of the scatter plot of ellipses are noteworthy:

1. First, there is a rather strong negative relationship between parental education “gaps” and ethnicity “gaps.” That is, states with a high degree of social inequality tend to also exhibit a high degree of ethnic inequality. New York is a case in point; lying in the upper left quadrant, New York has a “parental education gap” of about 0.30 points (half a standard deviation) and an “ethnicity gap” of around 0.60 (a full standard deviation).
2. Some degree of inequality is present in nearly all states. This inference is based on noticing that nearly the entire scatter of ellipses lies above 0 on the vertical axis (indicating positive parental education effects within states) and below 0 on the horizontal axis (indicating that African American ethnicity is associated with lower levels of disciplinary climate).
3. However, the magnitude of inequality varies quite substantially across states. There is a cluster of states near the origin (the point indicating equality on both parental education and ethnicity). There are also states far from the origin (e.g., New York, New Jersey, California, and Massachusetts), implying substantial inequality in access to favorable disciplinary climate in these states.

[8] The mean differences associated with social inequality are adjusted for ethnicity, and the mean differences associated with ethnicity are adjusted for parent education. The 95 percent confidence ellipses are based on the empirical Bayes posterior distribution (Morris 1983) of the parental education and ethnicity coefficients for each state.

Figure 8. 95 Percent Bivariate Confidence Ellipses for the State-specific Coefficients Associated with Parental Education and African American Ethnicity (Outcome: Mean School Climate)



4. There is considerable overlap among the ellipses, making it hard to distinguish many pairs of states and, in fact, making pairwise comparisons confusing. However, the ellipses of any pair of states can be shaded (as Michigan's ellipse in figure 8) to facilitate a desired pairwise comparison. Using computer graphics, it is easy to highlight any subset of states to generate clearer comparisons.

The value of the ellipses is that they automatically communicate the degree of uncertainty about rankings among states. Consider, for example, Michigan and Ohio. Ohio is characterized by significantly greater ethnic inequality than Michigan is, i.e., the gap between European Americans and African Americans in the disciplinary climates they encounter is statistically greater in Ohio than in Michigan, as indicated by the fact that the two ellipses do not overlap on the horizontal axis. However, the two states do not differ in social inequality, as indicated by the fact that their ellipses do overlap on the vertical axis.

Excellence versus Equality

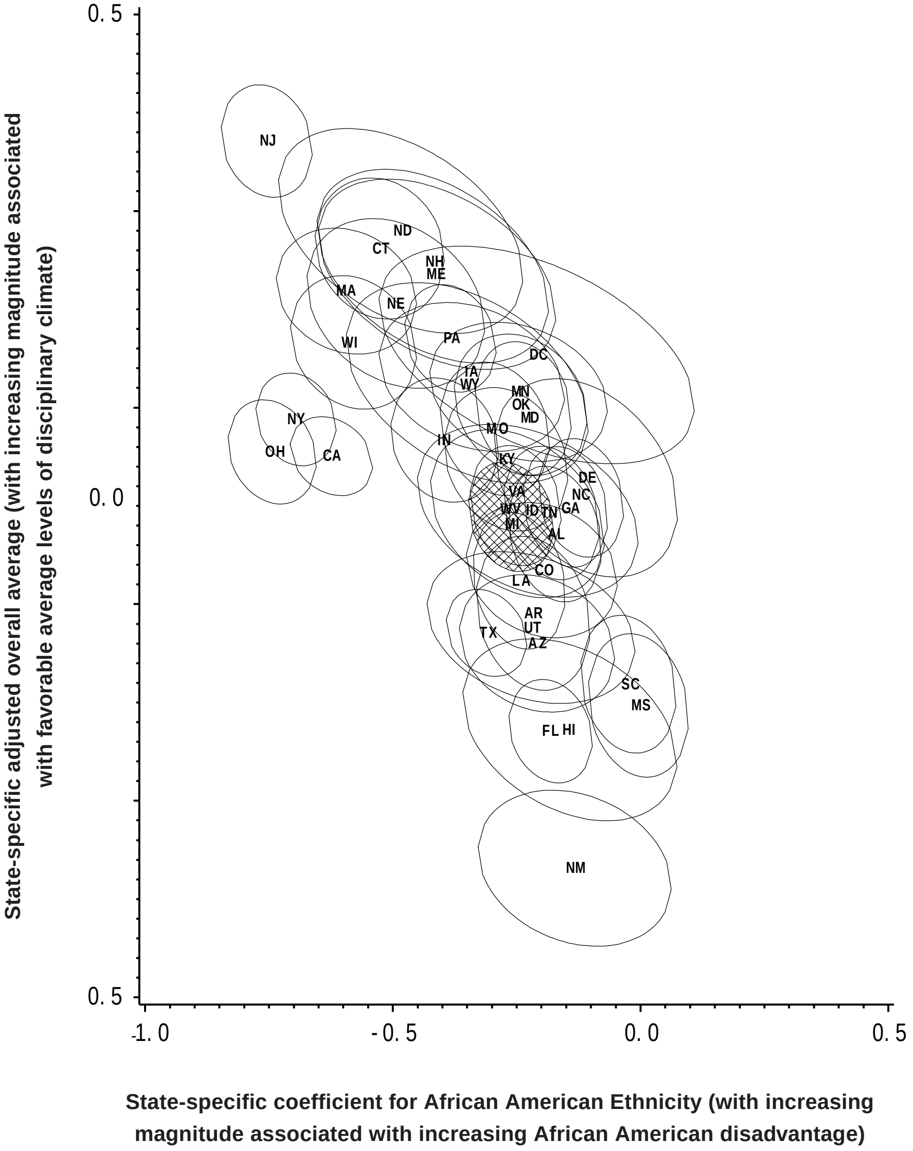
It is also possible to plot "excellence" (high levels of a resource) against "equality," as depicted in figure 9. The figure shows, for example, that New Jersey, though displaying a comparatively high degree of ethnic inequality, has one of the highest average levels of disciplinary climate. Equality is not a good thing if environments are equally bad; South Carolina and Mississippi exhibit low levels of inequality but also low average levels of disciplinary climate.

For the other resources, the pooled results also poorly represent the degree of inequality in some states. Again, the data reveal substantial evidence of state variation. It is possible and generally useful to describe state-to-state variation in access to these resources as we did in the case of school climate (figures 4 and 5). However, a detailed discussion of differences among the 41 states on all resources goes beyond the scope of this paper.

Conclusions

The Trial State Assessment of NAEP reports mean student proficiency in a given subject for each of the participating states, broken down by ethnicity and parental education (c.f., Mullis et al. 1993). Although reports of state means are essential as part of an assessment of the condition of education in the U.S., we have argued in this paper that such state means, by themselves, are difficult

Figure 9. 95 Percent Bivariate Confidence Ellipses for the State-specific Coefficients Associated with Intercept and African American Ethnicity (Outcome: Mean School Climate)



to interpret and even misleading. The means reflect an unknown mix of contributions from student demographics, school organization and process, and state policy. To supplement the reporting of means, we have proposed a reporting of the access that states provide to key resources for learning. Knowing the extent to which states provide these resources to students of varied social background and ethnicity points toward sharply defined policy debates concerning ways to improve education. The results of our analysis are both substantive and methodological.

Substantive Findings

Our results indicate substantial inequality in access to resources, on average, over the 41 participating states. Social background, as indicated by levels of parental education, is significantly related to access to a school with a favorable disciplinary climate and a school that offers high school algebra for eighth graders. Social background also predicts the probability that an eighth grader will encounter a teacher who majored in mathematics and a teacher who emphasizes reasoning during mathematics instruction. These effects of social background are adjusted for ethnicity.

The results for ethnicity parallel those for social background, though they vary to some degree by the resource of interest. For example, with respect to school disciplinary climate, European Americans encounter, on average, the most favorable disciplinary climates; Asian Americans and Native Americans are next, followed by Hispanic Americans and finally by African Americans. The probability of attending a school that offers algebra is distributed a little differently: Asian Americans experience the highest probability of attending such a school; European Americans, African Americans, and Hispanic Americans are next most likely to attend such a school; and Native Americans have the lowest probability of attending such a school. These effects of ethnicity are adjusted for social background. The results for teacher preparation and emphasis on reasoning are more complex: ethnic gaps in access are greatest at highest levels of parental education, with Asian Americans and European Americans having greater access than other groups to each resource.

In sum, we have found substantial evidence of inequality in access to these resources as a function of social background and ethnicity. However, there is also substantial variation across states in the extent of inequality. While some degree of both forms of inequality appears to exist in nearly all states, inequality is much more pronounced in some states than in others. Moreover, the overall level of availability of each resource also varies from state to state. While a fine-grained analysis of state differences on all four resources would be of interest, such a study goes beyond the scope of the current paper. However, we have suggested ways in which state differences might be examined.

The policy implications of these findings vary as a function of the resource in question. Whether a school offers algebra to eighth graders is amenable to direct influence by state and district policy. The key impediment to offering algebra in a given setting is cost. It is generally more costly for smaller schools than for larger schools to diversify their curricula. Similarly, hiring teachers with serious college-level preparation in mathematics is under the direct control of policy, with cost again being a key impediment.

Constructing a favorable disciplinary climate, in contrast, is only partially under the control of policymakers. Effective adult leadership in a school setting is arguably the primary ingredient in creating such a climate, though the active participation of students and parents is also required for success. Skill, knowledge, and commitment are required, and there is considerable uncertainty about how to foster the needed efforts. Similarly, a decision to emphasize reasoning is in the hands of the teacher, depending on the teacher's knowledge, skills, and evaluation of student needs. Interventions to encourage instruction that emphasizes reasoning are currently widespread, but the outcomes of such interventions are inevitably uncertain.

In sum, how information from a report such as ours ought to influence the policy debate will vary as a function of the kind of resource in question. Options for increasing access to certain resources must be evaluated in terms of cost and feasibility. Our primary point, however, is that systematically collected data on access to key resources, as a supplement to reports of mean proficiency, ought to constitute an important input into policy debates regarding educational reform.

Methodological Implications

The educational resources considered here clearly constitute a small subset of those that ought to be studied. We have reasoned that the resources of key interest are those suggested by prior theory and research and operationalized in NAEP. There should also be some evidence that the NAEP indicator of the resource relates as expected to key educational outcomes. The logic of this argument is to extend NAEP to include a wider range of possible resources than are now included and to take some pains to insure that the resource indicators achieve a modicum of construct validity. For example, it would be extremely useful to field-test and validate student reports of multiple indicators of student social background including parental occupation, and to construct and validate a scale for cognitive stimulation in the classroom based on student reports. Linking NAEP data to indicators of neighborhood demographic characteristics such as poverty concentration, housing density, and ethnic composition would strengthen inference by allowing control for residential context. And it would be exciting to include with NAEP a survey of teachers in order to construct school-level indicators, based on teacher reports, of normative cohesion, expectations, collaboration, control, opportunities for learning, and school-level academic press. The availability of denser data at the level of the student, classroom, and school would provide a wider range of school resources than can now be studied, leading to a richer characterization of the association between student background and access to resources.

A promising avenue for future research is to develop more sophisticated models to explain variation in access to key resources. School district wealth, urban versus suburban versus rural location, school size, per pupil expenditures, and school social composition may shape the probability that resources will become available to a student; and studying such predictors may shed light on impediments to increasing access and identify new targets for intervention by policy. Our broad recommendation is that, as we assess student progress in subject-matter proficiency, we also assess the extent to which the education system provides resources that support such student progress.

References

Arnold, C. L., Kaufman, P. D., and Sedlacek, D. S. (1992). *School Effects on Educational Achievement in Mathematics and Science: 1985–1986 (Research and Development Report)*. U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1996). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Boston, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Berne, R. (1994). Educational Input and Outcome Inequities in New York State. In R. Berne and L. O. Picus (Eds.), *Outcome Equity in Education* (pp. 1–23). Thousand Oaks, CA: Corwin Press, Inc.

Bernstein, B. (1970). Education Cannot Compensate for Society. *New Society 387*: 344–347.

Bowles, S., and Gintis, H. (1976). *Schooling in Capitalist America*. New York: Basic Books.

Breslow, N., and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association 8:* 9–25.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods.* Newbury Park: Sage Publications.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1994). *An Introduction to HLM: Computer Program and Users' Guide. Version 2*. Chicago: Department of Education, University of Chicago.

Bryk, A. S., and Thum,Y. M. (1989). The Effects of High School Organization on Dropping Out: An Exploratory Investigation. *American Educational Research Journal 26*(3): 353–383.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Finley, M. K. (1984). Teachers and Tracking in a Comprehensive High School. *Sociology of Education 5*: 233–243.

Fraser, E. (1959). *Home Environment and the School.* London: University of London Press.

Fuller, B. (1987). Raising School Quality in Developing Countries: What Investments Improve School Quality? *Review of Educational Research 57*: 255–291.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association 85*: 398–409.

Glass, G. V. (1976). Primary, Secondary, and Meta-analysis of Research. *Educational Researcher 5*: 3–8.

Johnson, E. G., Mazzeo, J., and Kline, D. L. (1993). *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kozol, J. (1991). *Savage Inequalities*. New York: Crown.

Lee, V. E., and Bryk, A. S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education 62*: 172–192.

Lee, V. E., and Bryk, A. S. (1993). The Organization of Effective Secondary Schools. In L.Darling-Hammond (Ed.), *Review of Research in Education* (Ch. 5, pp. 171–267). Washington DC: American Educational Research Association.

Little, R. J., and Schenker, N. (1994). Missing Data. In G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 39–75). New York: Plenum Press.

Medrich, E., and Griffith, J. (1992). *International Mathematics and Science Assessments: What Have We Learned.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Mislevy, R. J. (1992). Scaling Procedures in NAEP. Special Issue: National Assessment of Educational Progress. *Journal of Educational Statistics 17*(2): 131–154.

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association 78*(381): 47-65.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993, April). *NAEP 1992 Mathematics Report Card for the Nation and the States.* Princeton, NJ: Educational Testing Service.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*. New Haven, CT: Yale University Press.

Oakes, J. (1990). *Multiplying Inequalities: The Effects of Race, Social Class, and Ability Grouping on Access to Science and Mathematics Education*. Santa Monica, CA: RAND.

Page, R. N. (1990). The Lower Track Curriculum in a College-preparatory High School, *Curriculum Inquiry 20:* 249–282.

Peaker, G. F. (1967). *The Plowdon Children Four Years Later*. London: National Foundation for Education in England and Wales.

Raudenbush, S. W., and Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics 20*(4): 307–335.

Raudenbush, S. W. (1994). Random Effects Models. In H. Cooper and L. Hedges (Eds.), *The Handbook of Research Synthesis* (Chapter 20, pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W. (1988). Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics 13*: 85–116.

Raudenbush, S. W., Fotiu, R. P., and Cheong, Y. F. (in press). Synthesizing Results from the Trial State Assessment. *Journal of Educational and Behavioral Statistics.*

Raudenbush, S. W., Fotiu, R. P., and Cheong, Y. F. (1998). Inequality of Access to Educational Opportunity: A National Report Card for Eighth Grade Math. *Educational Evaluation and Policy Analysis 20*(4): 253–268.

Raudenbush, S. W., Rowan, B., and Cheong, Y. F. (1993). Higher Order Instructional Goals in Secondary Schools: Class, Teacher, and School Influences. *American Educational Research Journal 30*(3): 523–553.

Rist, R. (1970, August). Student Social Class and Teacher Expectations: The Self-ful filling Prophecy in Ghetto Education. *Harvard Educational Review 40*: 411–451.

Rosenbaum, J. E. (1976). *Making Inequality: The Hidden Curriculum of High School Tracking*. New York: Wiley.

Rosenthal, R., and Jacobson, L. (1968). *Pygmalion in the Classroom*. New York: Holt, Rinehart and Winston.

Rotberg, I. C. (1998, May 15). Interpretation of International Test Score Comparisons. *Science 280 (S366):* 1030–1031.

Rowan, B., Raudenbush, S. W., and Cheong, Y. F. (1993). Teaching as a Non-routine Task: Implications for the Organizational Design of Schools. *Educational Administration Quarterly 29*(4): 479–500.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., and Smith, A. (1979). *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*. Cambridge, MA: Harvard University Press.

Ryan, W. (1971). *Blaming the Victim*. New York: Pantheon.

U.S. Department of Education. National Center for Education Statistics. (1995). *The Condition of Education 1995*. Washington, DC: U.S. Government Printing Office.

Willms, J. (1992). *Monitoring School Performance: A Guide for Educators*. Washington, DC: Falmer Press.

Wolf, R. (1968). The Measurement of Environments. In A. Anastasi (Ed.), *Educational Testing Problems in Perspective*. Washington, DC: American Council on Education.

Zeger, S., Liang, K., and Albert, P. (1988). Models for Longitudinal Data: A Likelihood Approach. *Biometrics 44*: 1049–60.

Moving Educational Research Toward Scientific Consensus

David W. Grissmer
Ann Flanagan
RAND

Introduction

Educational research has been characterized, perhaps unfairly in recent years, by the inconsistency of its research results and by a lack of consensus across its broad and multidisciplinary research community (Wilson and Davis 1994; Saranson 1990). The broad purpose of this conference is to help determine how we can improve the consistency and accuracy of results in educational research so that we can build a base of knowledge widely accepted by this diverse research community and, more importantly, by teachers, principals, superintendents, and policymakers. To do so is a daunting task since education is one of the most complex topics addressed by social science. It is not surprising that progress in this direction has been slow, given both the broad interdisciplinary basis and the inherent complexity of learning.

We have proceeded with the hope that better nonexperimental data and more sophisticated model specifications and estimation techniques will eventually bring consensus. In this paper we will suggest that simply improving the kinds of nonexperimental data currently collected, along with the associated statistical methodologies, will never be sufficient to achieve the kind of scientific consensus needed to effectively guide educational policies.[1] Research shows that the effects we are trying to measure are quite complex. They often appear to be nonlinear, sensitive to contextual factors, moderately correlated among themselves, and subject to selection bias within families and schools. More-

[1] Support for this work came from the Center for Research on Educational Diversity and Excellence (CREDE), the NAEP Redesign Research Program, the NAEP Secondary Analysis Program, and Exxon Corporation.

over, some achievement effects are long-term, sustained long after an intervention has stopped; and some fade after a few years. These results may be only the tip of the iceberg, considering the complexity of the underlying developmental phenomena we are trying to understand.

This complexity places great demand on the quality of our data, the sophistication of our model specifications, and the accuracy of our estimation techniques. One interpretation of the wide variation in measurements of the effects of most factors affecting student achievement is that our data, model specifications, and estimation techniques do not yet reflect much of this inherent complexity. When results vary, it is difficult to determine why one set of results should be trusted over another, since practically every measurement makes different assumptions or uses different model specifications and estimation techniques. The wide variety of data quality, assumptions, and specifications may introduce enough bias and randomness to produce inconsistent effects across different data sets and model specifications. In this case the results should not be interpreted as “no effect,” but rather as inconclusive.

We suggest that three research approaches will be necessary to lead reliably to research consensus: increasing experimentation, building theories of educational process, and improving our nonexperimental analysis. Further, we believe that future data collection and research should be guided by a strategic plan built upon experimentation. Such a plan would provide the necessary data to build theories of educational process and improve our specifications of models used in nonexperimental analysis.

Experiments—if well designed, implemented, analyzed, and replicated—provide explanations that are as close to causal as possible in social science. Such experiments can provide the most accurate results for the effect of a particular variable in a given context. Experiments can also play another, and perhaps more important, role in social science research—namely, helping to validate model specifications for nonexperimental data. A key theme of this paper is that future experimentation and data collection need to be directed toward both the building of theories and the improvement of our assumptions in analyzing nonexperimental data. In the long run, policy analyses will largely be dependent on improving nonexperimental analysis since experiments can never be counted on to solve all the complex and contextual effects present in education.[2] Therefore, improving our confidence in the model specifications used with nonexperimental data is critical.

The major thrust of this paper is to suggest that building scientific consensus will require a coherent research strategy. This strategy must be built upon increasing experimentation, developing theories of educational process, and improving confidence in nonexperimental analysis, if we are to achieve research consensus. In this paper we focus initially on the broad lack of agreement relating to the effects of educational resources and social and educational policy on children. Thirty years of research with nonexperimental data have led to almost no consensus on these important policy issues. We then focus on a narrower question, namely, the impact of resources on educational outcomes, particularly student achievement. This situation presents an interesting case study where a consensus based on the results of nonexperimental data once existed, only to be challenged recently by new experimental and nonexperimental research.

We use the Tennessee class size experiment results to illustrate the process of deriving "rules" for model specification used in nonexperimental data involving class size. We then illustrate the process of building theories of educational process related to class size effects and describe the role of such theories in building stronger consensus. Finally, we specifically focus on implications for the National Assessment of Educational Progress (NAEP) and other data collections and more generally suggest directions for future research and development (R&D) efforts to build a more solid foundation of knowledge for educational policymaking.

Children's Well-Being: The Ongoing Debate

Federal, state, and local governments spend approximately $500 billion per year in social, educational, and criminal justice expenditures on the nation's

[2] Large-scale experiments such as the Tennessee class size experiment can be costly and take considerable time to plan, implement, and analyze. While more experimentation seems essential to making progress in educational research, educational research will probably never follow health research, where trials are needed for every new intervention before implementation.

children and youth (Office of Science and Technology Policy 1997).[3] The amount spent on children appears to have increased substantially over time (Fuchs and Rekliss 1992), although there is debate about the magnitude of the real increase in spending. Thus, an important set of public policy questions is associated with how effective this increased spending has been at improving the well-being of our children. Besides increased investment, there have been significant changes in families, communities, and schools that would be expected to affect children's outcomes.

There is little scholarly consensus about the effects of expenditures on children or the effects from changing families, communities, and schools. For instance, scholars disagree about the impact of the War on Poverty and expanded social welfare programs (Herrnstein and Murray 1994; Jencks 1992); they also disagree on whether increased school resources have raised student achievement levels (Burtless 1996; Ladd 1996a). There is disagreement about the way communities have changed for black families (Wilson 1987; Jencks 1992) and whether the net effect on children of recent changes in the family has been positive or negative (Cherlin 1988; Zill and Rogers 1988; Fuchs and Rekliss 1992; Popenoe 1993; Stacey 1993; Haveman and Wolfe 1994, 1995; Grissmer et al. 1994). There is more agreement about the effects of desegregation, although some dispute remains (Wells and Crain 1994; Schofield 1995; Armor 1995; Orfield and Eaton 1996). Finally, many small-scale, intensive early childhood programs appear to produce significant short- and long-term effects, but there is disagreement about large-scale programs—how large the effects from attending kindergarten and preschool are and how long these effects last (Barnett 1995; Karweit 1989). Recent evidence suggests that the cost-effectiveness of early childhood programs can depend critically on the characteristics of the targeted group, with significant net fiscal returns for some groups, but not others (Karoly et al. 1998).

[3] This estimate does not include the foregone taxes for deductions for children and day care. Besides public sector spending on children, approximately $560 billion is spent in the private sector on children, bringing the average public and private spending per child to approximately $15,000 annually. This amount is estimated assuming the cost of raising a child to age 18 to be approximately $150,000, with approximately 70 million individuals between the ages 0–18. Thus, annual expenditures are $150,000 x 70,000,000/18 = $560 billion. See United States Department of Agriculture (1997) for estimates of the cost of raising children.

Despite the lack of consensus among the educational research community, dramatic changes are being proposed and are occurring in both social and educational policies, based on perceptions that past policies have failed. For instance, much of the movement toward more fundamental reform of public schools arises from perceptions that massive increases in resources in grades K–12 education over the last 25 years have resulted in declining—or at best stable—student achievement (as measured by scores on the Scholastic Achievement Test [SAT] and NAEP scores) and that schools have particularly failed minority students. If so, a solid case could be made for restructuring school governance and incentive structures so that more effective utilization of resources might possibly occur (Hanushek 1994; Hanushek and Jorgenson 1996). However, new research is challenging this once widely accepted conclusion.

A Shifting Consensus: The Effects of Educational Resources[4]

Until the early to mid-1990s, the dominant research position among social scientists was that school resources had little impact on student achievement. This counterintuitive view dated from the "Coleman report" (Coleman et al. 1966). Influential reviews by Eric Hanushek (1989, 1994, 1996, 1999) also argued that evidence from over 300 empirical measurements provided no consistent evidence that increases in school resources raised achievement scores. It was suggested that a key reason for inefficiency in public schools was a lack of incentives (Hanushek and Jorgenson 1996).

However, it would not be surprising that some money was spent inefficiently, given that no definitive results emerged from educational research that could guide policymakers. At worst—if past resources can be shown to have had no effect on achievement—this finding can simply indicate the lack of guidance by good R&D. The lack of a critical level of R&D funding and critical mass of high quality research may provide an explanation for inefficiency just as persuasive as the lack of incentives (Wilson and Davis 1994).

[4] The early sections of this paper draw heavily from four recent papers—Grissmer, Flanagan, and Williamson (1998a); Grissmer et al. (1998); and Grissmer, Flanagan, and Williamson (1998b); and Grissmer et al. (forthcoming). We have quoted liberally from these papers without quotation marks.

Hanushek's original reviews did not group studies using the quality of data and specifications, type of intervention, or student or grade level (1989, 1994). However, Hanushek refined his reviews, focusing on effects from per pupil expenditure and pupil/teacher ratio reductions and disaggregating studies by grade level, level of aggregation, and model specifications (Hanushek 1996, 1999). These later reviews still indicated that subsets of studies provide positive and negative coefficients in about equal numbers. One focus was on studies using a production function framework where the previous year's test scores were used as controls. These models were judged by many to be the most likely to avoid bias. These models also showed balanced numbers of positive and negative coefficients. These results strengthened the conclusion that the nonexperimental evidence supported little effect from class size reductions or additional expenditures.

Subsequent literature reviews questioned the selection criteria used in Hanushek's reviews to choose studies for inclusion and the assignment of equal weight to all measurements from the included studies. Two subsequent literature reviews (Hedges, Laine, and Greenwald 1994; Krueger 1999a) used the same studies included in Hanushek's reviews, but came to different conclusions. One study used meta-analytic statistical techniques for combining the measurements, which do not weigh each measurement equally (Hedges, Laine, and Greenwald 1994). Explicit statistical tests were made for several variables for the hypotheses that the results support a mean positive coefficient and reject a mean negative coefficient. The results concluded that, for most resource variables, the results supported a positive relationship between resources and outcomes. In particular, per pupil expenditures and teacher experience provided the most consistent positive effects, with pupil/teacher ratio, teacher salary and teacher education having much weaker effects.

A more recent literature review using the same studies included in Hanushek's reviews also concludes that a positive relationship exists between resources and outcomes (Krueger 1999a). This review criticizes the inclusion and equal weighting of multiple measurements from single published studies. Some studies provided as many as 24 separate measurements due to the presentation of sets of results for many subgroups. Since the average sample size will decline as subgroups increase, many of the measurements lacked the statistical power to detect policy-significant effects; and thus many insignificant coefficients might be expected. Since the presentation of results for subgroups is not done uniformly across studies, and may even be dependent on the results obtained, Krueger (1999a) reanalyzes the data to determine if the inclusion of multiple measurements significantly affects the conclusions reached. His analysis concludes that the inclusion of multiple measurements is a significant factor in explaining the original conclusions, and that less weight placed on these multiple measurements would lead to support for a positive relationship between higher per pupil expenditures and lower pupil/teacher ratio and outcomes.

A more comprehensive review of the literature prior to 1990 used meta-analytic statistical comparison techniques, but searched a wider literature and imposed different quality controls (Greenwald, Hedges, and Laine 1996). All the included studies used achievement as the dependent variable and measurements at the individual or school level only. The resulting set of measurements utilized in the study included many measurements that were not included in Hanushek's studies and rejection of about two-thirds of the measurements included in Hanushek's reviews.

The conclusions analyzing the set of coefficients from six variables (per pupil expenditure, teacher ability, teacher education, teacher experience, pupil/teacher ratio, school size) supported statistically the hypothesis that the median coefficients from previous studies showed positive relationships between resource variables and achievement. However, the variance in coefficients for each variable across studies was very large. Extreme outliers appeared to be a problem for some variables, and the coefficients across studies appeared to have little central tendency indicating the presence of nonrandom errors.

This review also reported results for measurements using different model specifications (longitudinal, quasi-longitudinal and cross-sectional).[5] The results showed that median coefficients changed dramatically for most variables across specifications, with no recognizable pattern. Although few studies had what were considered to have superior specifications (longitudinal studies), the median coefficients for these models were negative for per pupil expenditure, teacher education, pupil/teacher ratio, and school size. When the median coefficients of studies having quasi-longitudinal studies were compared to coefficients from the entire sample, results were similar for four variables, but differed for the remaining two variables by factors ranging from 2 to 20. In the

[5] Longitudinal studies were defined as those having a pretest control score, and quasi-longitudinal was defined as having some earlier performance-based measure as a control. Cross-sectional studies merely had SES-type variables included as controls.

case of teacher salary, these studies provided a median coefficient indicating that a $1,000 salary increase could boost achievement by over one-half standard deviation.

This review utilized better screening criteria and better statistical tests to conclude that the overall evidence supported positive effects from additional resources. However, the large variance in coefficients and the sensitivity of the median coefficients to which studies were included provided little confidence that the literature could be used to estimate reliable coefficients. In particular, models thought to have superior specifications provided no more consistent results and sometimes provided noncredible estimates.

Besides the argument from literature reviews, Hanushek made another argument that seemed consistent with his conclusions. Measured in constant dollars, expenditures per pupil doubled between the late 1960s and the early 1990s; however, NAEP scores at age 9, 13, and 17 showed no dramatic improvement in average reading or math skills during this period. We address this argument next.

Interpreting NAEP Score Trends

Achievement scores are a particularly good measure of the changing environment for our children since research has shown that achievement reflects the combined influence of families, communities, and schools. Significant changes in the quality of our families, schools, and communities should be reflected on achievement trends that are best measured by NAEP (Cambell et al. 1996; Miller, Nelson, and Naifeh 1995; Mullis et al. 1993; Reese et al. 1997).

The NAEP achievement scores collected from 9-, 13-, and 17-year-olds since 1969 are the only nationally representative achievement scores available. The primary purpose of NAEP has been to simply monitor the achievement of American students; however, NAEP scores are increasingly being used to evaluate the effects on youth from the dramatic changes in families, communities, and schools, and from our nation's educational and social policies—changes that have taken place since the late 1960s. These changes include the following:

- National efforts to equalize opportunity and reduce poverty that began in the mid-1960s and continued or expanded in subsequent decades. These efforts included federally funded preschools (e.g., Head Start), compensatory funding of elementary schools with large numbers of low-income students, desegregation of schools, affirmative action in college and professional school admissions, and expanded social welfare programs for poor families.
- Changes in school attendance and school changes that were not primarily designed to equalize opportunity. These changes included increased early schooling, greater per pupil expenditures, smaller classes, significant changes in the characteristics of teachers, and systemic reform initiatives.
- Changes in families and communities that may have been somewhat influenced by efforts to equalize opportunity and reduce poverty but that occurred mainly for other reasons. Specifically, parents acquired more formal education, more children lived with only one parent, more children had only one or two siblings, and the proportion of children living in poverty rose. At the same time, poor blacks concentrated more in inner cities, while the more affluent blacks moved to the suburbs.

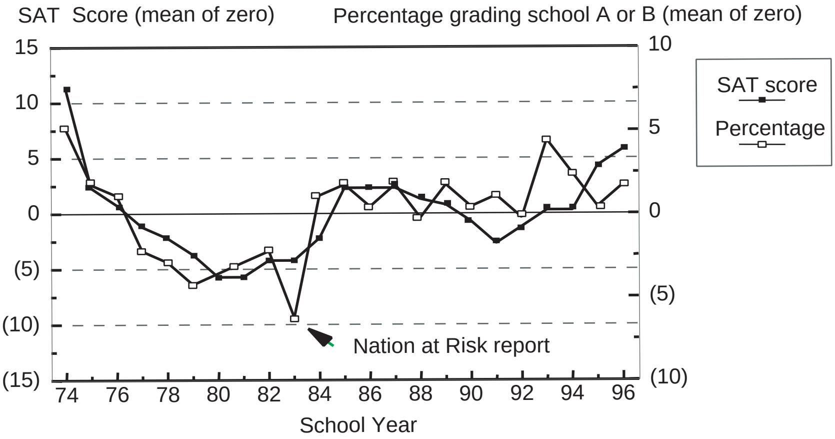
The 17-year-olds tested by NAEP in 1971 would have grown up in families and communities and attended schools largely unaffected by the changes cited above. However, those recently tested would have lived their entire lives in families, communities, and schools reshaped by these policies. It would be hard to take a position about the quality of our families, communities, and schools and the effectiveness of social and educational policies that would be inconsistent with the trends in the NAEP data.

Until recently, the NAEP scores were used only peripherally to address these kinds of questions, partly because the more widely recognized (but fatally flawed) SAT scores were used whenever test scores entered the public debate. One reason that SAT scores are used effectively in public debate is that the public appears to base its assessment of the quality of American schools on SAT scores (Grissmer forthcoming). Figure 1 shows the results of an annual public opinion poll that asks adults to grade the nation's schools. The percentage of adults giving schools an "A" or a "B" is graphed against changes in annual average SAT scores.[6] The data show that public opinion appears to follow the SAT trends.

The well-known flaws in the SAT scores for monitoring national achievement trends result from their self-selected sample (Advisory Group on the Scholastic Aptitude Test Score Decline 1977; Koretz 1986, 1987; Rock 1987; Grissmer et al. 1994). The scores are biased downward, not only because of an increasing percentage of students taking the test but also because the students making the largest achievement gains from 1970 to 1990—minority and disadvantaged students—are largely missed by the SAT because they do not go to college. Ironically, if K–12 education improves, allowing more children to attend college, the SAT scores will decline. Thus, SAT scores are probably a perverse indicator of K–12 school quality.

The research community switched to analyzing NAEP data in isolated studies dating from the mid-1980s. A steady stream of analyses from the late 1980s drawn from the NAEP data developed into more detailed analyses using

Figure 1. Comparing the trends in SAT scores with percentage of adults giving schools a grade of "A" or "B"



[6] The graph normalizes both variables to a mean of 0. The regression fit for the equation, School grade = a + b (Average SAT score), gives b = .79 (t = 5.2), R-Squared = .56.

new methodologies from the mid-1990s.[7] Early work took note of the large gains in black scores and the very small gains in white scores, along with the resulting convergence of the black-white test score gap. The contrast with falling SAT scores was noted. However, familiarity with this earlier work—buttressed by the National Research Council (1989)—seemed to remain confined to a small group of researchers, and declining SAT scores remained the dominant influence among both the public and the research community.[8]

Starting in the early 1990s, analyses of the NAEP data began to provide more detail about differences in trends among black, Hispanic, and white students; differences in trends for lower- and higher-scoring students; differences by age; and particularly differences by entry cohorts. The analyses also attempted to explain the trends and the convergence in the black-white test score gap.

Across ages and subjects, the largest gains in scores occurred for black students; but significant gains were registered by Hispanic students and lower-scoring white students, with small gains or none registered by average and higher-scoring white students (Hedges and Nowell 1998; Hauser 1998; Grissmer et al. 1994, 1998; Grissmer, Flanagan, and Williamson 1998a). These studies also noted the evidence that black gains were largely confined to a group of about 10 cohorts born in the mid-1960s to the mid-1970s and entering school around 1970 to 1980. For later cohorts, black scores and the black-white achievement gap have—for most age groups and subjects—remained stable or declined.

The most striking feature of the NAEP results for blacks is the size of adolescents' gains for cohorts entering from 1968–1972 to 1976–1980. These

[7] See Hauser (1998) for a history of utilizing NAEP scores from 1984 to 1992. This period included work by Jones (1984); Koretz (1986, 1987); National Research Council (1989); Linn and Dunbar (1990); and Smith and O'Day (1991). See Rothstein (1998) for a long-term history of achievement that extends through 1997. This paper draws from all of these studies.

[8] This phenomenon points to a second problem in attaining consensus in the educational research community. While small groups of researchers with in-depth knowledge in a subject may find consensus, it is quite another problem for this information to be disseminated, accepted broadly, and commonly cited in most research. The diverse set of journals and disciplinary boundaries make it difficult for narrow consensus to become broad consensus.

gains were 0.6 standard deviation averaged across reading and math. Such large gains for very large national populations over such short time periods are rare, if not unprecedented. Scores on IQ tests given to national populations seem to have increased gradually and persistently throughout the 20th century, both in the United States and elsewhere (Flynn 1987; Neisser 1998). But no evidence exists in these data involving large populations showing gains even close to the magnitude of the gains made by black student cohorts over a 10-year period.

Even in intensive programs explicitly aimed at raising test scores, it is unusual to obtain gains of this magnitude. Early childhood interventions are widely thought to have the largest potential effect on academic achievement, partly because of their influence on brain development. Yet only a handful of "model" programs have reported gains as large as half a standard deviation (Barnett 1995). These programs were very small-scale programs with intensive levels of intervention. Even when early childhood programs produce large initial gains, the effects usually fade at later ages. Among blacks who entered school between roughly 1968 and 1978, in contrast, the gains were very large among older students and were not confined to small samples, but occurred nationwide.

Beginning in the mid-1990s, finding the likely causes of these gains became the focus of research. Part of the quest was to determine whether the dramatic changes that occurred in families during this period could explain the gains. Utilizing data from several sources (Current Population Survey [CPS], the National Longitudinal Survey of Youth [NLSY], and the National Education Longitudinal Study [NELS]), one study developed a new methodology to estimate the size of the net expected gains from changes in eight key family characteristics for 13- to 17-year-old test-takers from 1970–90 (Grissmer et al. 1994). The analysis required several assumptions—one concerning the stability of family coefficients in achievement equations over time.[9] The results of the analysis indicated that changes in the family would predict small positive gains in scores for all racial-ethnic groups and that these gains could account for the smaller score gains among whites but could explain only about one-quarter of the minority gains.

[9] Evidence from Hedges and Nowell (1998) and Cook and Evans (1997) appears to support fairly stable family coefficients over time.

Another analysis using NAEP individual level data also concluded that family effects could account for only one-quarter or so of black gains (Cook and Evans 1997). This analysis relied on student-reported family characteristics collected with the NAEP, but utilized a methodology newly imported from labor economics to attempt to partition the gains into those related to family changes, changes in family structural characteristics, and those due to changes between and within schools. If effects from changing family characteristics are small, the likely remaining hypothesis for the black score gains is school-related, community-related, or related to yet unmeasured family characteristics.

Jencks and Phillips (1998) summarized research efforts focusing on the black-white test score gap. Their book brought together a diverse set of scholars to try to determine where consensus can be achieved on this topic and where and what kind of additional research is needed.[10] Three analyses reported in the book look at the convergence and possible divergence of the black-white score gap for cohorts born as early as 1950 (Hedges and Nowell 1998; Phillips, Crouse, and Ralph 1998; Grissmer, Flanagan, and Williamson 1998a). Two of the studies utilize NAEP data as well as achievement and survey data from other studies. All agree that significant narrowing occurred for cohorts born prior to about 1978—but no further narrowing occurred for later cohorts.

Although the black-white gap for reading actually widened, Phillips, Crouse, and Ralph (1998) concluded that the widening is not statistically significant. Hedges and Nowell (1998) and Grissmer, Flanagan, and Williamson (1998a) provided evidence that family changes may explain a part of the narrowing. Further, Grissmer, Flanagan, and Williamson (1998a) observed that the timing of the black gains by age group and region suggested two major hypotheses for the gains. The first hypothesis was based on changes in schooling—changing pupil/teacher ratios and class sizes, changing teacher characteristics, and changing curricula. Changing pupil/teacher ratios emerged

[10] In the process of achieving consensus, support for a continuing series of books dedicated entirely to exploring the most important questions in education seems crucial. Besides Jencks and Phillips (1998), Ladd (1996a) and Burtless (1996) are also good examples. In these latter books, the consensus might be characterized more by what is not known than what is known.

as a viable, but not completely satisfactory, explanation in other analyses (Krueger 1998; Ferguson 1998).[11]

A second explanation emerged, more closely related to the changes engendered by the Civil Rights movement and the War on Poverty. Such changes could have direct effects related to school desegregation—particularly in the South—and indirect effects caused by the perceived shift in the motivation for and attitudes toward education of black parents and students stemming from better opportunities for future schooling and jobs. An additional possible shift from these efforts could have occurred in the behavior and attitudes of teachers of black students that resulted in increased attention and resources. The timing of the black gains by age coincides with the broad-scale implementation of such efforts, if the assumption is made that most of the effects would occur only if students experienced these changes from the early grades forward. The large gains for minority and disadvantaged students, as well as the smaller gains (or lack of gain) among average and higher-scoring white students, pose a challenge to the thesis that the increased spending in education and social programs aimed at these students was ineffective.

Analysis of NAEP scores appears to be central to the debates about changes in American families and schools, policies providing equal opportunity in education, and the best way to spend investments in education and children. The effective absence of these scores from these national debates has allowed many widespread beliefs to proliferate that seem to be at odds with the NAEP results. The NAEP data do not suggest that families have deteriorated since 1970. Nor do they suggest that schools have spent money inefficiently or that social and educational policies aimed at helping minorities have failed.

Instead, they suggest that family environments changed in positive ways from 1970 to 1996, that the implementation of the policies associated with the Civil Rights movement and the War on Poverty may be a viable explanation for large gains in black scores, and that certain changes in our schools and curriculum are consistent with NAEP score gains. While the NAEP scores

[11] The timing of pupil/teacher ratio changes would suggest that score gains should have started earlier and would affect white scores as well—leading to overpredicted white gains. Further research to determine whether class size for black students fell more than for white students might help reduce the overprediction of white score gains. This overprediction would also be addressed if class size reductions were small or nonexistent for more advantaged white students.

alone cannot reject the beliefs about deteriorating families and schools and the ineffectiveness of social and educational policies, the advocates of such beliefs must provide an explanation for NAEP scores consistent with their positions. The NAEP scores from 1971 to 1988 generally support a more positive picture of our families, schools, and public policies; however, trends in black achievement since 1988 to 1990 have been more discouraging, and it is critical to understand why these reversals have occurred.

Trends in School Resources

Research on NAEP scores shows that the increases were negligible only for the higher-scoring white population, but substantial for black, Hispanic, and lower-scoring white students. A second line of research using new data and new methods of estimating "real" per pupil expenditures over time shows that resource growth tended to occur where achievement gains were made (Rothstein and Miles 1995).

A new method of deflating school expenditures, taking account of the labor intensity of schools, showed that resources did not come close to doubling as had been indicated by the commonly used Consumer Price Index (CPI). Use of more appropriate indices for adjustment of educational expenditures reflecting their labor intensity provides much lower estimates of real growth (Rothstein and Miles 1995; Ladd 1996b).

Moreover, the new method—developed to assign school expenditures to programmatic categories that could distinguish spending on different types of students—showed that even this smaller increase overestimates the additional resources available to boost achievement scores for regular students. A large part of the smaller estimated increase went for students with learning disabilities, many of whom are not tested.[12] Another part also went for other socially

[12] There is agreement that a disproportionate fraction of the expenditure increase during the NAEP period was directed toward special education (Lankford and Wyckoff 1996; Hanushek and Rivkin 1997). Hanushek and Rivkin estimated that about a third of the increase between 1980 and 1990 was related to special education. NAEP typically excludes about 5 percent of students who have serious learning disabilities. However, special education counts increased from about 8 percent of all students in 1976–77 to about 12 percent in 1993–94. These figures imply that 7 percent of students taking the NAEP tests were receiving special education resources in 1994, compared to 3 percent in 1976–77. This percentage is too small to have much effect on NAEP trends, but it should in principle have had a small positive effect.

desirable objectives that are only indirectly related to academic achievement. Taking into account better cost indices, and including only the spending that would have been directed at increasing achievement scores, Rothstein and Miles (1995) concluded that the real increase in per pupil spending on regular students was closer to 30 than to 100 percent.

These smaller additional expenditures for regular students are mainly accounted for by lower pupil/teacher ratios, increased teacher salaries due to more experienced and educated teachers, and compensatory programs that would be expected to benefit minority and lower income students (Rothstein and Miles 1995; Hanushek and Rivkin 1997). The key issue then becomes whether these resource increases can plausibly explain any part of the pattern of large black gains and the absence of white gains unaccounted for by family changes. This pattern might be explained if black students received disproportionate shares of the additional resources or if black students benefited more than white students due to similar increases in resources.[13]

The Tennessee Experiment

Important new evidence for challenging the view that money doesn't matter comes from a large-scale experiment in Tennessee on the effects of class size. The Tennessee experiment in education was largely ignored for several years by the wider research community, and only recently has been reanalyzed and given its deserved prominence (Ritter and Boruch 1999). This experimental research suggests that reductions in class size may, in fact, have more impact on disadvantaged and minority students than on white students. A quasi-experiment in Wisconsin that varied student/teacher ratio also provided new evidence (Molnar et al. 1999).

The first experimental evidence on the effect of major educational variables came from a Tennessee study on the effects of class size (Word, Johnston, and Bain 1990; Finn and Achilles 1990; Mosteller 1995). About 79 schools in

[13] A number of policies sought to shift resources toward minority or low-income students during these years, including federal compensatory funding based on the percentage of children in poverty, school desegregation, and court-directed or legislative changes in state funding formulas toward minority and low-income school districts. However, other factors operated over this time period that could have increased funding for middle- and upper-income children as well. It is still unclear whether the net effect has been to disproportionately shift resources toward minority and lower-income children.

Tennessee randomly assigned about 6,000 kindergarten students to class sizes of approximately 15 or 23 students, and largely maintained their class size through third grade. Additional students entering each school at first, second, and third grade were also randomly assigned to these classes making the entire experimental sample approximately 12,000. After third grade, all students were returned to standard, large-size classes through eighth grade. The students in the experiment were disproportionately minority and disadvantaged—33 percent were minority, and over 50 percent were eligible for free lunch.

Analysis of the experimental data shows statistically significant, positive effects from smaller classes at the end of each grade from K–8 in every subject tested (Finn and Achilles 1999; Krueger 1999b; Nye, Hedges, and Konstantopoulos 1999; Nye, Hedges, and Konstantopoulos forthcoming). The magnitude of results varies depending on student characteristics and the number of grades in small classes. Measurement of effect sizes from four years in small classes at third grade varies from 0.25 to 0.4 standard deviation (Krueger 1999b; Nye, Hedges, and Konstantopoulos forthcoming). The current measurement of long-term effects at eighth grade show sustained effects of approximately 0.4 standard deviation for those in small classes all four years, but little sustained effect for those in smaller classes one or two years (Nye, Hedges, and Konstantopoulos 1999). Short-term effects are significantly larger for black students and somewhat larger for those receiving free lunches.[14]

Questions were raised whether the inevitable departures from experimental design that occur in implementing the experiment biased the results (Krueger 1999b; Hanushek 1999). These problems included attrition from the samples, leakage of students between small and large classes, possible nonrandomness of teacher assignments, and schooling effects. Recent analysis has addressed these problems without finding any significant bias in the results (Krueger 1999b; Nye, Hedges, and Konstantopoulos 1999; Nye, Hedges, and Konstantopoulos forthcoming; Grissmer 1999). It is possible for further analysis to find a flaw in the experiment that significantly affects the results, but extensive analysis to date has eliminated most of the potential problems.

[14] Long-term effects have not been reported by student characteristics. Following the experiment, Tennessee also cut class sizes to about 14 students per class in 17 school districts with the lowest family income. Comparisons with other districts and within districts before and after the change showed even larger gains of 0.35 to 0.5 standard deviations (Word, Johnston, and Bain 1994); Mosteller 1995). Thus the evidence here suggests that class size effects may grow for the most disadvantaged students.

The Wisconsin SAGE (Student Achievement Guarantee in Education) quasi-experimental study differed in several important ways from the Tennessee STAR experiment (Molnar et al. 1999). In the SAGE study, only schools with very high proportions of free-lunch students were eligible for inclusion. Assignments were not randomized within schools, but rather a preselected control group of students from different schools was matched as a group to the students in treatment schools. The treatment is more accurately characterized as pupil/teacher ratio reduction since a significant number of schools chose two teachers in a large class rather than one teacher in a small class. The size of the reduction in pupil/teacher ratio was slightly larger than the class size reductions in Tennessee.

There were about 1,600 students in the small pupil/teacher treatment group in Wisconsin, compared to approximately 2,000 students in small classes in Tennessee. However, the size of control groups differed markedly—around 1,300 students in Wisconsin and around 4,000 in Tennessee, if both regular and regular-with-aide classes are combined. The SAGE sample had approximately 50 percent minority students with almost 70 percent eligible for free or reduced price lunch.

The results from the Wisconsin study for two consecutive first grade classes show statistically significant effects on achievement in all subjects (Molnar et al. 1999). The effect sizes in the first grade are in the range of 0.1–0.3 standard deviations. The lower estimates between 0.1–0.2 occur in regression estimates, while the raw effects and hierarchical linear modeling (HLM) estimates are in the 0.2–0.3 range. While the estimates seem consistent with the Tennessee study at first grade, more analysis is needed before the results can be compared.

Learning From the Tennessee Experiment about Model Specification

One of the problems with nonexperimental data analysis is that the research community usually fails to completely list the assumptions that are required in any analysis to make the analysis equivalent to experimental data.[15] Such listing of assumptions would make much more explicit the wide gap that exists between experimental and nonexperimental data analysis.

Partly because there have been so few experiments in education, we have not paid much attention to their potentially critical role in shaping theories about education, helping to correctly specify variables and models using nonexperimental data, and specifying what data we should collect. If applied to reliable experimental data, models used to estimate nonexperimental data should be able to duplicate the experimental results. Krueger (1999b) suggests that production functions with previous year's score do not duplicate the Tennessee effects except in the first year of smaller classes. This larger first-year effect has been interpreted as a socialization effect.

The Tennessee results suggest several further specification issues. First, schooling variables in one grade can influence achievement at *all* later grades, so conditions in all previous years of schooling need to be present in specifications. Second, a pretest score cannot control for previous schooling characteristics. *The Tennessee results suggest that two students can have similar pretest scores, similar schooling conditions during a grade, and emerge with different posttest grades influenced by different earlier schooling conditions.* For instance, despite having similar schooling conditions in grades 4–8, relative changes in achievement occurred in those grades for those having one to two or three to four years in small classes in K–3. Another way of stating this analytically is that effect sizes at a given grade can depend on interactions between this year's schooling characteristics and all previous years' characteristics.

The production function framework using pretest controls assumes that any differences in pre- and posttests are captured by changed inputs during the period. The Tennessee results suggest that coefficients of such specifications are *un-interpretable from a policy perspective* since the effect of a change in resources during a period cannot fully be known until past and future school-

[15] An excellent counterexample is Ferguson and Ladd (1996), which starts to describe the conditions for a "gold standard" model and provides one of the most complete listings of assumptions of any economic analysis. Raudenbush and Wilms (1995) and Raudenbush (1994) also carefully outline the statistical assumptions in two kinds of models used in education. See also Heckman, Layne-Farrar, and Todd (1996) for an analysis that tests and provides evidence of the weakness of the assumptions inherent in a certain kind of model linking educational outcomes to educational resources.

ing conditions are specified. Thus the answer to the question of whether a smaller class size in second grade had an effect cannot be known until later grades, and the answer will depend on what the class sizes were in previous and higher grades.

Another interpretation of the Tennessee data is possible—namely, that reduced class size is a multiyear effect whose precise pattern is dependent on duration. Being in a small class not only raises short-term achievement in the current year, but also has an effect in succeeding years. Then the effect in first grade consists of a residual effect from kindergarten plus an independent first grade effect. The second grade effect is the sum of the residuals from kindergarten and first grade, plus an independent second grade effect. This explanation would account for the increasing effect with more years in small classes in the K–3 years—but would also account for the pattern after return to larger classes after third grade. Clearly there is residual, and continuing, effect from having attended smaller classes in grades K–3. However, the permanence of the effect depends on duration, indicating the effects are not simply additive.

Conceptually this makes the effect of class size reductions resemble a human "capital" input that can change output over all future periods, and models specifying the effects of capital investments may be more appropriate.[16] Production functions generally assume constant levels of capital, but children's human "capital" is probably constantly changing and growing.

From the standpoint of child development, these results are consistent with the concepts of risk and resiliency in children (Masten 1994; Rutter 1988). Children carry different levels of risk and resiliency into a given grade that appear to interact with the schooling conditions in that grade to produce gains or losses. For instance, four years of small classes appear to provide resiliency against later larger class sizes, whereas one year or two years do not.

Few, if any, previous studies have included variables for prior years' school characteristics from elementary school. At the individual level, virtually no longitudinal data from kindergarten were available. At more aggregate district

[16] Production functions are typically applied to model complete growth cycles in agriculture or other areas. We have tried to apply it to much smaller increments of growth in children by using pre- and post-test results. Production functions may have done less well in earlier studies predicting weekly plant growth as oppused to the complete cycle of growth over a season.

and state levels, data are usually available describing average characteristics for earlier years, but were probably seldom used.

Since most data sets at the individual level, such as NELS, do not contain the previous year's history for grades K–8, they cannot be used to estimate class size effects under this hypothesis.[17] Probably most previous measurements at the individual level have not had such data, and this might explain the downward bias in results. However, models using aggregate data have more of a chance at being able to include previous history—on average—for students in the sample. For instance, at a school district level, data would be available on class sizes in previous years. If no in-migration and out-migration occurs, then the average class size for district students can be determined for previous years. Migration will weaken the validity of these estimates, which means that higher levels of aggregation (state level data) will likely capture more accurately the historical class size for students in the aggregate sample.

The usual tendency for researchers is to trust the results of individual level analysis more than those of aggregate level analysis. This trust arises from several factors: larger sample size, more variance in variables, and sometimes more detailed family data. However, individual level analysis is to be preferred over aggregate level only if the quality of variables is equivalent. If aggregate level data can better capture accurate historical information, then these estimates may produce better results. Another implication is that our data collection efforts should focus on longitudinal data from early years. Use of longitudinal data beginning at or prior to school entry can sort out some of the specification problems that may exist in previous analyses.

There are two new sources of such longitudinal data that will include school, teacher, and family characteristics and achievement data. First, there are the newly emerging longitudinal state databases that link student achievement across years. Such data have very large sample sizes, and linkages are possible with teacher data and school characteristics. These data will be better able to address some of the potential specification issues involving dependence

[17] The current year's class size will work if it is highly correlated with all past years' class sizes. However, at the individual level it seems likely that the random elements that determine year-to-year class size—including in-migration and out-migration and decisions when to create additional classes—would not make this year's class size a particularly good predictor of previous years' sizes, particularly over many grades. However, this correlation should be explored.

of later achievement on the previous year’s class size as well as thresholds and interactions with teacher characteristics. It may also be possible to determine class size effects in later grades as well as in early grades. The second source will be the Early Childhood Longitudinal Study (ECLS) funded by the U.S. Department of Education, which will collect very detailed data on children, their families, and their schools. These data will be much richer in variables, but much smaller in sample size than the state data sets.

A Weak Test of the Hypothesis Using State NAEP Data

Analysis of state NAEP scores is providing preliminary supportive evidence that certain state policies *do* matter in improving scores, that minority and disadvantaged students show the most gain from increased resources, and that the distribution of key resources is inequitable (Grissmer et al. forthcoming; see also Raudenbush in this volume).

We have used the state NAEP data for the seven reading and math tests given between 1990 and 1996 at the fourth or eighth grade level to test two hypotheses:

- whether aggregate state results provide estimates of pupil/teacher ratio that are in reasonable agreement with the Tennessee class size effects; and
- whether these results change when we utilize a pupil/teacher ratio variable incorporating only the current year of the NAEP test vs. the average of all previous years in school.

Estimates have been made using the 271 average state scores in equations controlling for the effects of different family and demographic characteristics of students across states (Grissmer et al. forthcoming). We have utilized three different ways of controlling for family characteristics at the state level. We have supplemented the NAEP family characteristics with Census data to derive more accurate family variables than those provided by NAEP (Grissmer et al. forthcoming). We have also utilized SES-like variables derived from the NELS and Census data. We found little difference in results across these family measures.

We used a random-effects model and estimated with the generalized linear estimator with exchangeable correlation structure, which takes account of the lack of independence of state observations across tests (produces robust standard errors), the unbalanced panels, and heteroskedascity. We also have made estimates with generalized least squares and maximum likelihood, achieving almost identical results.

In the equations linking average state scores to family and state educational characteristics, we included four educational variables that account for 95 percent of the variance in per pupil spending across states. These variables are average teacher salary, pupil/teacher ratio, teacher-reported adequacy of resources, and percentage of students in a state in public prekindergarten.[18] We found the expected signs and statistical significance for pupil/teacher ratio, teacher-reported resources, and prekindergarten participation. We found insignificant results for teacher salary.

The pupil/teacher ratio effect in this model would predict a rise of about 0.14 standard deviation for reduction of eight pupils per class (approximately the size of the Tennessee class size reductions). This effect is markedly smaller than the reported Tennessee class size effect of around 0.20–0.25. However, if we include in our models an interaction term allowing larger pupil/teacher effects for states with more disadvantaged students, we find markedly larger effects for states having more disadvantaged students. The Tennessee experimental sample contained a disproportionate percentage of minority and free lunch students, compared to all Tennessee students (Krueger 1999b). If we take into account the characteristics of the Tennessee sample and the interaction effect, the equations would predict a class size effect for the Tennessee sample that agrees with the actual effect.

We have tested whether results for pupil/teacher ratio differed in our data set when the variables were defined using pupil/teacher averages during time in school vs. pupil/teacher value in the year of the test only. We use the state average pupil/teacher ratio during all years in school, the average during grades 1 through 4, and the value in the year of the test. The estimates for these vari-

[18] We used a pupil/teacher variable rather than class size, since data were only available by year by state for the pupil/teacher ratio. While the two are highly correlated, one cannot necessarily assume that reductions in pupil/teacher ratio and class size would produce the same effects.

Table 1. Comparing Three Pupil-Teacher Coefficients That Incorporate Differing Information about Previous Grades

	Random effect		Fixed effect	
Variable	**coef**	**t-value**	**coef**	**t-value**
Average P/T during school years	-0.015	-2.41	-.014	-1.16
Average P/T in grades 1-4	-.020	-2.69	-.026	-2.60
P/T in year of the test	-.008	-1.32	.014	1.57

ables are shown in table 1 for random and fixed effect models. The results show that including current year pupil/teacher ratio instead of information from previous years causes the coefficients generally to weaken in both random and fixed effect models and to change signs in one model.

The Investments That Do Matter

The long debate about the role of resources in education has finally shifted from whether money does matter to what kinds of investments do matter for what kinds of children. The earlier conclusions drawn from reviews of the nonexperimental literature (Hanushek 1994)—that money has not mattered due to the inefficiency of our public school system and its lack of incentives—appear flawed. Over the last 25 years, money invested in schools for regular education students has gone mainly to develop programs targeted at minority and disadvantaged youth, lower pupil/teacher ratios, and raise average teacher salaries. Evidence is emerging that at least two of these investments have paid off for minority and disadvantaged students—lowering pupil/teacher ratios and targeting resources to minority and disadvantaged children. However, at least part of the money used to reduce pupil/teacher ratio for students from families with higher SES levels—the majority of students—may have been spent inefficiently.

Still, the broad-ranging conclusions that money does not matter in education without substantial changes in the existing structure of and incentives in public education are contradicted by experimental evidence and the results presented here. Moreover, the evidence supporting these conclusions now appears to be based on poor model specifications. This leaves the more viable hypothesis—that money *does* matter if invested in the right programs and targeted toward minority and disadvantaged students (Grissmer, Flanagan, and Williamson 1998b; Grissmer 1999).

Implications for Future Methodology and Data Collection

We suggest several specific ways that data and methodology might be improved, as follows: building micro-level models of educational processes, conducting more experiments, and improving NAEP data. For the latter, we discuss such measures as using school district samples rather than school samples, collecting additional family variables, improving children's responses (especially with regard to reporting levels of parental education), collecting additional information from teachers, using state Census data to improve the individual level variables, using supplementary data from the Census, and collecting additional parent information.

Building Micro-level Models of Educational Processes

The results of either experimental or nonexperimental analysis are meant to provide the material for developing theories of educational processes and student learning that gradually incorporate wider phenomena in their purview. Eventually, these theories should accurately predict the results of empirical work and be able to make new predictions to guide future empirical work. Theories by their very nature are more robust than any set of experimental or nonexperimental studies since they incorporate results of multiple measurements and incorporate research across levels of aggregation. However, little theory building has been done in education.

Hierarchies exist in science whereby certain areas of science are derived from and built upon the knowledge in more basic science. For example, the science of chemistry relies partly upon basic knowledge in physics for explanation. The science of biology is partly built from knowledge of chemistry; and, within biology, molecular biology provides some basis for the applied science of medicine. Typically the ordering of these hierarchies is derived from the size of the basic building blocks studied. Physics studies elementary particles and atoms. Chemistry studies combinations of atoms. Biology studies complex combinations of atoms with certain structures (genes, etc).

Education is far up in the hierarchies of social science. It rests upon knowledge derived from psychology, cognitive and brain science, genetics, sociology, child development, psychopathology, and economics. It is one of the more complex "sciences" that depends on good basic science in the lower hierarchies. Without linking the knowledge from these more basic sciences, educational research will never have a solid foundation. Educational research needs to incorporate the findings of these more basic sciences in building its theories, data collections, and methodologies. An example is the need to understand why smaller class sizes seem to produce higher levels of student achievement and why the results are multiyear and can be either short- or long-term.

Research directed toward measuring class size effects has generally treated the classroom as a black box in which only inputs and outputs are needed and in which knowledge of the transforming processes inside are unimportant for purposes of measurement. The current analytical methods also isolate the cause and effects of class size reductions within precise time periods in a way that seems at odds with the more continuous, cumulative, and often delayed effects that occur in children's cognitive development. Reconciling the differences in experimental and nonexperimental evidence will probably require a far better understanding of the underlying mechanisms occurring in classrooms and the developmental process in students that determine achievement.

In the case of class size, we need a theory of classroom and home behavior of teachers, students, and parents that answers why smaller classes might produce higher achievement in both the short and the long term. Initially we need to understand what teachers and students do differently in large and small classes and then whether these differences can be related to the size of short-term achievement. Perhaps the more difficult area of theory will be to explain gains long after the end of an intervention. An early intervention either has to change cognitive, psychological, or social development in important ways or change the future environment (e.g., peers, families) that affects the individual. Possibilities range from changes in brain development to learning different ways of interaction with teachers and peers to developing different study habits to being in different peer groups years later.

Answering these types of questions not only requires different types of data collection, but also requires understanding much about psychology, child development, and individual behavior (teacher and student). We provide some simple examples in the appendix at the end of this paper of the types of modeling and data collection that spring from alternate hypotheses about why smaller class sizes work.

One type of theory-building would use *time on task* as a central organizing concept in learning. A secondary concept involves the productivity and optimal division of that time among the different alternatives: new material through lectures, supervised and unsupervised practice, periodic repetition, and review and testing.[19] Students have a wide variance in the ways they spend time in school and at home, and it is likely that home time can substitute for specific types of teacher time.

Some research suggests that significant differences may exist in the amount of instructional time and the ways in which it gets used across different types of classes and different teachers and by students with different characteristics (Molnar et al. 1999; Betts and Shkolnik 1999a; Rice 1999). A theory of learning needs to be developed that incorporates school and home time and the various tradeoffs and differences that exist across teachers, classrooms, and SES levels. Such a theory would generate a number of testable hypotheses for research, which would then allow better and probably more complex theories to be developed. Such theories would then provide guidance as to what research is important to undertake.

Such theory-building would mandate linking several disparate and isolated fields of research in education. There is micro-research involving time on task, repetition, and review in learning specific tasks. There is research on teachers in classrooms. There is research on homework and tutoring. There is research on specific reading and math instructional techniques. There is research on class size and teacher characteristics. Theorists can begin to understand these disparate areas and suggest theories that can explain the empirical work across these areas. Such linkages seem essential to future progress.

Finally, cognitive development may have patterns of development similar to other areas of development in children, since brain development seems to be central to each type of development. There is much research on patterns of physical, emotional, and social development in children from birth, differences across children, delays in development, and dependence on previous mastery. Studies involving long-term developmental outcomes—especially for children at risk—identify resiliency factors that enable development to oc-

[19] This approach is best exemplified in Betts and Shkolnik (1999a, 1999b) and Betts (1997).

cur even in highly risky situations. Much can be learned from this literature that can help prevent researchers from making poor modeling assumptions.

Need for Experiments

A major question raised by many other researchers, and currently under discussion, is the role of experimentation in educational research and other areas of social science (Burtless 1993; Boruch and Foley 1998; Boruch 1997; Hanushek 1994; Heckman and Smith 1995; Ladd 1996a; Jencks and Phillips 1998). Many interesting and complex issues arise in thinking about future experimentation, but consensus is emerging on the need for *more* experimentation in education.

Certainly, the value of the Tennessee experiment suggests that a selected number of social experiments may considerably add to our consensus knowledge in education. Besides the accuracy of the direct results, experiments tell us how to get more reliable results from nonexperimental data. Although expensive to carry out, experiments may be cheap compared to the costs of ineffective educational policies.

However, experimentation is much easier in smaller settings than in the classic, large-scale social experiments such as that produced in Tennessee. A very simple set of experiments could be designed around classroom- and school-level variables that would be much easier to carry out, yet could provide a better underlying base of information on which to build educational theories. For instance, simple experiments that divide children who miss a particular test question into two remediation groups with retesting could help locate the cause of missed questions and help develop efficient methods of remediation.

Improving the NAEP Data

The NAEP data are becoming so central to issues in both educational and social policy that priority should be given to significant expansion and improvement. We address two issues with respect to the NAEP data: (1) redesign of the sample to be district- rather than school-based and (2) improving family variables.

A School District NAEP Sample

The hypothesis suggested here implies that the lack of historical data on schooling variables may prove to be a barrier to unbiased results with individual level NAEP data. Here we focus on one option that would improve the aggregate data analysis possible with NAEP data. If NAEP could become a school district sample rather than a school sample, then historical data from school districts (not available at the school level of aggregation) could be used in the formulation of variables.

A district level sample would also result in improved family variables in NAEP data, since Census data would be available for most school districts. Currently, family variables in NAEP cannot be improved with Census data at the school level because privacy concerns prohibit their use within school areas. A school district sample would also address another NAEP deficiency—namely, the absence of several educational policy variables not available at the school level, such as per pupil spending. A much wider and better defined set of educational policy variables is readily available at the school district level and is already collected. Thus, a school district, rather than school level, NAEP sample would be desirable from the standpoint of improving family controls and educational policy variables.

A straightforward random sample of students at the district level would involve additional administrative costs, because the districtwide student universe would be needed and administration of tests would have to occur across many schools or involve assembling students from many schools in a central location. Such a sample would also have the disadvantage that, while Census and educational policy data would be available at the district level, certain school level characteristics obtained from student data at the school level would be missing. For instance, the school level sample of students is often used to define the characteristics of peers and their families. So a trade-off would occur with a district sample in that the educational and family characteristics would improve, but less would be known about some of the local, school level characteristics. Much of this missing school level data could probably be collected using enrollment data available at the school level. For instance, instead of using the sample of 20 students per school to estimate percentage minority, this figure would be obtained from schoolwide enrollment data.

Another change that would occur with a district sample would be that the sample of teachers surveyed would increase substantially. Currently, a typical classroom sample is 10–25 students, and a single teacher survey is collected. In a district sample, there would be few students selected from the same classroom, so the teacher sample would approach more closely the size of the student sample. The larger teacher sample would have some advantages besides increased size. The desired teacher variables are the characteristics of all teachers of the students from the time they entered school. The current teacher sample of one to two teachers per school, since it is a very small sample, is a very weak proxy for the characteristics of teachers at the school or the characteristics of all previous teachers of the students. Obtaining a much larger sample of teachers at the district level would provide a better proxy for the kinds of teachers likely to have taught in the district.

It may be possible to combine school level and district sampling to obtain a reasonable sample for each. About one-half of public school districts have fewer than 1,000 students and only one or two elementary schools per district. Thus, this sample of school districts would be close to the size of a school sample. However, these districts constitute only about 6 percent of total students. At the other end of the spectrum, there are about 300 districts with over 20,000 students, which account for nearly one-third of all students. In these districts, the number of schools ranges from about 30 to over 600. In most of these districts, a district sample could be drawn based on samples of schools, with 5–10 students per school. The remaining 60 percent of students are in school districts where some limited clustering by school could occur, but a sound district sample would probably have to include students from most schools.

However, it may be feasible to design a joint district- and school-based sample that samples fewer students per school. Such a sample would have several analytical advantages. It would contain an additional hierarchy in the sample—the district level, where extensive and better data exist on families and schools. It could still contain school-based samples, but with fewer students per school. It would also enlarge the number of teachers surveyed. Such a sample design would, however, entail additional costs since more schools would be sampled, district samples would require more effort at developing universe files, and more teachers would be surveyed.

The question is whether the analytical advantage would be worth the additional cost. To answer this question, we suggest a two-stage feasibility analysis in which a preliminary assessment by a group of statisticians and researchers would be performed to see whether serious barriers exist, to develop preliminary cost estimates, and to better define the analytical advantage. This group would either recommend a more detailed study and assessment or make the judgment that the analytical advantage is probably not worth the cost.

One of the chief advantages of moving to a district sample is that comparisons of scores could be made for major urban and suburban area school districts. It is the urban school systems that pose the largest challenge to improving student achievement, and being able to develop models of NAEP scores across the major urban school districts could provide critical information in evaluating effective policies across urban districts. The sample sizes would be much larger than at the state level and could be expected to provide more reliable results than for states.

Improving Family-level Variables

The primary objective of NAEP has always been seen as monitoring trends in achievement rather than explaining those trends. One result of this philosophy is that few family variables have been collected with NAEP. Compared with family data collected with other national achievement data or on other government surveys dealing with children's issues, NAEP collects very few family variables. In addition, the quality of the family variables collected has always been questioned since they are reported by the students tested. The perception of weak family variables may partially explain why NAEP scores have not been utilized more frequently in research on educational and social policies.

We have compared the accuracy of NAEP family data with Census data at the state level and analyzed the sensitivity of our estimates with state NAEP data with NAEP variables, Census variables, and SES variables formulated from parent-reported NELS data (Grissmer et al. 1998). Not surprisingly, we find that NAEP variables for race and family type (single-parent or two-parent) match Census data well, once differences in the samples are accounted for. However, students substantially inflate their parents' education level at the college level. Fourth graders report 58 percent of their families include a college graduate compared to 26 reported in the Census; comparable figures for eighth graders are 42 percent compared to 25 percent in the Census. However, reports of "high school only" and "not a high school graduate" are much more accurate. Students appear to be unable to distinguish between "some college" and "college graduate"—and individuals using NAEP data should combine these two categories when using the data.

There are several ways that the family variables can be improved in the NAEP data collection. We describe six increasingly complex options.

Collecting additional variables from children. There are two variables that are strongly significant in equations linking family characteristics and achievement that can be easily included and that children probably could report with some accuracy. The first variable is family size (number of siblings), which should present little problem for student reporting. The second is current age of mother. The age of mother combined with the child's age would enable the variable of age of mother at birth to be computed. Some pretesting may be required to determine the method of asking these questions, but even reporting mother's age in gross categories—five-year groupings—would be an improvement.

Recent research is finding that two-parent families with a stepparent do not have similar effects as do two biological parents (McLanahan and Sandefur 1994). The effects on children from a family including a stepparent appear to be closer to single-parent effects than to living with two biological parents. So information that could distinguish two-parent biological families from those with a stepparent would be useful. Adding a question on whether the parents are divorced is one approach. Asking separate questions about living with each parent is another approach.

One other variable that should be considered is locus of control. Locus of control is derived from a set of questions focusing on the perceived ability to affect life events. There are now more specific sets of questions that focus on specific events or conditions such as school performance. Locus of control has been collected in the NELS and NLSY data sets and is strongly statistically significant in equations relating achievement to family characteristics after all the common family characteristics are entered.

Improving children's responses. It appears that students have the least knowledge about post–high school education levels of parents. One hypothesis is that children have simply never asked parents about education level. Another is that parents report inaccurate levels of education to children, somewhat inflating their own level of education. In the former case, it may be possible to have children formally or informally ask parents prior to the test. This could take the form of a simple request before the test or a more formal written form for the parents to fill out. Pretesting this approach could help determine which hypothesis is causing the inaccuracy in reporting.

Collecting supplemental data from teachers. While individual level parental characteristics are desirable, teachers of NAEP students currently fill out an extensive survey that could be used to obtain family information. Teachers currently do not provide information concerning the socioeconomic characteristics of their students. Teachers could be asked several questions concerning the characteristics of the groups of students in their classes that might improve the data on family characteristics. These questions would take the form of identifying percentages of the students that fall into various categories. Income levels would probably be the most useful information. Giving teachers broad categories of income could prove better than the category of free and reduced price lunch as a control for family income. Items could include estimates for nearly all the important family variables. Such information could be first collected on a trial basis at low additional cost, perhaps for one year and utilized to see whether it improves the models.

Using state census data to improve individual level variables. We have utilized Census data to improve NAEP family variables at the state level. If NAEP data were only to be analyzed at the state level, the Census data combined with NAEP data could probably provide good estimates of all family background variables. However, the real value of NAEP data lies in the individual level data, and direct Census data have not been available at that level. So similar techniques cannot be used to directly derive school or individual level Census estimates.

It is possible to improve some of the reported NAEP variables at the school and individual levels by using the knowledge gained from state level comparisons. State level comparisons provide information about the accuracy of items such as parental education, and this information can in a limited way be used to impute better estimates to individual level variables. One simple application of this is to combine high school plus and college as a single category.

Further regressions across states linking the NAEP and Census estimates can provide information about how differences are connected to other family characteristics. For instance, the errors in reporting family education may be greater in states with high minority populations and lower incomes. This kind of information may be useful to impute better values at the individual level data. Such work would seek to better identify the types of students who report accurate and inaccurate data. However, while this approach should be tried, it would probably not result in dramatic improvements in the quality of individual level data.

Using supplementary data collection. The key information about family characteristics at the school level that would improve NAEP data might also be gathered directly from Census data. While privacy concerns limit the data available from Census at individual levels, the U.S. Department of Education would probably be able to obtain from Census data the school level population characteristics, if school boundaries were available. This would have to be done in conjunction with the NAEP data collection by collecting school boundary data on maps. Many school districts may be sufficiently large to allow Census to provide school data aggregated from the block level. This option should certainly be explored with the Census Bureau, and its cost assessed. There are many commercial vendors who can provide such data if given maps for specific disaggregated areas. The relative cost of this option compared to the cost of NAEP would be low.

The Census data could provide almost all the important background characteristics at the school level. But it would only be for all families in the area—not just the characteristics of families with fourth graders, for example. But the data would be highly correlated. Such data also could not track well the changes over time. Finally, the data would also be biased to the extent that the student population is not defined by specific geographical boundaries. But the advantages of this method would be the relatively low cost and the ability to provide a much richer set of characteristics at the school level.

Limiting parental data collection. Parental data collection for NAEP has always been a politically controversial issue, so extensive data collection similar to the type of collection performed on other U.S. Department of Education surveys is probably not feasible. The NELS, for instance, collects data from parents in an extensive survey. We consider here the minimum level of information which parents could provide that would enhance the NAEP data. The primary reason for parental data collection is to strengthen the individual level data in NAEP. A simple one-page form with no more than five items could solve the major problem with NAEP family data. It would take no more than a minute or two to fill out. It would ask for the key family background variables necessary for achievement score equations that are not accurately provided by the student. They include education level of each parent, family income in categories, and age of each parent. While a more extensive survey could certainly provide useful information, this minimum level of information would allow considerably more confidence in the use of individual level NAEP data without placing an undue burden on parents or children.

Summary

The interdisciplinary nature and the inherent complexity of educational research contribute their own set of challenges, but an additional reason for the lack of success in building consensus in educational research is the low investment in educational R&D and more broadly on R&D on children. On average, the nation spends approximately 2–3 percent of its gross domestic product for R&D. However, this proportion is not uniform across sectors of the economy, but can vary from less than 1 percent to approximately 20 percent (pharmaceuticals and integrated circuits) (Grissmer 1996). Currently, we spend less than 0.3 percent of educational expenditures for R&D, and less than 0.3 percent of expenditures for children are directed toward R&D on children (Consortium on Productivity in the Schools 1995, Office of Science and Technology Policy 1997). Compared to other sectors, this is a very low investment in R&D. Perhaps the reported problematical quality of educational R&D is partly due to the insufficiency of funding, when compared to its inherent complexity (Grissmer 1996; Wilson and Davis 1994; Atkinson and Jackson 1992; Saranson 1990). Alternately, the low funding level might reflect the poor quality of R&D.

Successful R&D is the engine that drives productivity improvement in every sector of our economy. Thus, strong R&D in education is a prerequisite to continual improvement in our education system and in our children's well–being. Without solid R&D, we will continue to go through wave after wave of reform without clearly separating the successful from the unsuccessful.[20] It is difficult to see how American K–12 education can become world class unless our educational R&D begins to build a more solid foundation of knowledge concerning education. If R&D can begin to play the role that it does in virtually every other sector of our economy, then continual educational improvement can be taken for granted, just as continual improvement in automobiles, computers, and life expectancy is now taken for granted.

[20] It is not that some reforms may not have been effective or had an impact on educational outcomes. The history of student achievement and educational outcomes suggests that scores have risen over long periods of time—and that students of a given era always seem to outscore their peers of earlier eras (Neisser 1998; Rothstein 1998). Rather, R&D could considerably improve the efficiency of the process of sorting the various reform initiatives and ensuring that the best are saved and the worst discarded.

Appendix

Simple Process Models of Class Size Effects

We start here by developing some simple models of the mechanism within classrooms that might cause class size effects and follow the implications of these assumptions on how we should specify models and why class size effects might be expected to have fairly wide variance. We do this simply to show that an important link is missing, a link that can guide us in specifying models and interpreting results of previous studies. If class size effects are produced by the kind of mechanisms assumed here, it implies that actual class size effects should have a wide variance and that some of the model specifications that were thought to be best actually can provide highly biased results.

Reductions in class size must change processes that occur in the classroom in order to have impacts on achievement. These differences in process that occur within smaller classes appear to determine whether class size affects achievement at all, whether effects are large or small, and whether effects widen or stay constant over several grades (Murnane and Levy 1996). In addition, the design of assessment instruments can determine whether class size effects are present in measurements.

Unless we know what processes change and how achievement is assessed, we cannot determine what model specifications and estimation techniques are appropriate. Since the data to determine what processes change in smaller class sizes are generally not collected, it will be difficult to sort out the reasons for the wide variance in the previous literature. We will discuss some simple, but extreme models to illustrate the point.

Demand for Teacher Individual Time

If we assume a “college professor” lecture model of classroom procedure, where there is essentially little or no interaction between teacher and student either during or after class and administrative time is borne by teaching assistants, then class size makes no difference. In this case, there is no cost to the teacher in having more students in the class. Class size makes a difference only when we assume that some teacher time is taken up by individual students—either through questions, special academic assistance, disciplinary actions, or administrative time (grading homework). In this case, additional students add to the teaching workload. If teachers have a fixed amount of time, then adding students can result in less time for presenting material or less time for student assistance. Thus, the size of the class size effect should depend on the portion of time teachers spend dealing with individual students (in one way or another) vs. time spent in general, lecture style instruction. In general, the more students need individual time, the larger will be the class size effect—other things being equal.

A second consideration is the variance among different types of students in requiring individual attention. A reasonable assumption here is that higher ability students or those with higher levels of family resources (broadly defined)—on average—will require less individualized attention. Essentially, substitution is occurring between family resources and school resources. In families with more resources, more of the students' academic and psychological needs are addressed at home, requiring less attention at school. This can include simple things such as helping with homework, enhancing learning opportunities, tutoring, and addressing the child's behavioral problems. For lower ability children or those with fewer family resources, more individualized attention will probably need to occur at school in order for them to achieve learning.

Thus, one would expect that class size effects would be larger for classes with lower ability students or students with fewer family resources. This also implies that there will be maximum class size levels (thresholds) that allow all the productive individual attention required, and above which no further class size effects will occur. But this threshold will vary by level of family resources.

Teacher and Curriculum Decisions

A third consideration is the teacher's reaction to scarcity of time. Teachers continually make choices about how fast to proceed with the scheduled curriculum, how much time to allow for slower students vs. faster students, and how much time to put into individual instruction vs. lecturing. With more students per class, these decisions become critical in determining whether class size effects occur. One scenario is that teachers slow down the pace of instruction in response to time scarcity. Individualized instruction is maintained for slower students, but less material is covered for all students. So the net effect is to cover less material for the school year for the whole class. Here, one might expect to see class size effects for the higher ability students (less material covered), but less so for those of lower ability.

Another teacher strategy is to cover all the material throughout the year (more time lecturing) and spend less time in individual instruction with slower children. Here, average class scores should shift downward in larger class size, but in a different pattern. Scores of higher scoring students would not be affected, but lower scoring students would have lower scores.

A crucial consideration in measurement is how the curriculum is adjusted the next year in response to these teaching strategies. If the effect of larger classes is failure to cover all the material for all students in the class, the next year's curriculum may or may not include all the material. If the curriculum accommodates this and starts where the previous year left off *for each student*, then over many years there will be an increasing gap between children in larger and smaller classes, i.e., the size of the effect will depend on the cumulative years in smaller class size. Thus, if smaller classes were instituted in grades K–8, one would expect to see a widening gap with each grade.

On the other hand, the start of next year's curriculum could begin *uniformly for all students* regardless of the amount of material covered last year. It could be started at the point where the larger or smaller class sizes left off. If it starts where the larger class sizes left off, then the gain from extra material covered in the smaller class in the previous year is lost. Thus, no cumulative effect is present, but a uniform score difference will be present each year. Essentially the smaller classes will cover additional material each year, but the gain from the previous year will be lost.

If the curriculum for all students is set where the smaller class sizes left off—leaving a permanent gap in coverage for those in larger class sizes—then whether the effect is cumulative depends on the extent to which mastery of the previous grade's material is required to perform well in the current grade. For subjects like math and reading, earlier mastery is probably more essential, and a widening gap would occur over several grades, i.e., the annual gap in material coverage would cumulate, causing further deterioration in later scores. On the other hand, in subjects like history or geography, where earlier mastery may be less important, the previous year's gap plays no role in next year's score, and a constant class size effect would be expected by grade.

Design of Assessment Instruments

Another consideration affecting the size of the measured class size is how assessment tests are designed. Designers of norm-referenced assessment tests first sample students' current knowledge at a given grade and develop a battery of questions that attempt to span the entire domain. A set of questions is chosen that provides a continuous range of question difficulty such that a different percentage of children answers each question correctly. Some questions nearly all students answer, while some are included that only a small percentage answer.

However, the domain of knowledge can depend on the size of classes attended by students. It is possible in some circumstances to have extra material covered by smaller class sizes included in assessments, while in other circumstances the extra material will not be part of the test. For instance, if tests were developed five years ago based on the then-existent domain of knowledge, and class sizes have declined since that time resulting in more material covered, the assessment instrument may not pick up the class size effect. Similarly, if assessment instruments are designed with students in larger classes prior to experimentation with smaller class sizes, then it is possible for the effects of class size to be attenuated if the instruments do not reflect possible additional material covered by smaller classes. In general, instruments designed to measure students' knowledge across several grades, rather than within each grade, and "re-normed" more frequently will be less vulnerable to these kinds of design effects.

Some Implications for Measurement and Specification

The above discussion illustrates that the size of the effect, its measurement, and its interpretation can depend on what occurs differently within the classroom when larger and smaller class sizes occur and on how assessments are designed. It implies that actual effects could vary considerably depending on different levels of student demand for individual time, teacher strategy, the coordination of the curriculum (e.g., year to year by class size), the different dependence by subject on previous knowledge, and assessment design. It would not be surprising in our decentralized educational system that smaller class sizes generate a wide variety of teacher and curriculum responses. Thus, ambiguity of results may not be surprising. Moreover, we may never be able to sort previous studies into groups with similar classroom process controls because the data along these kinds of dimensions were never collected for previous studies. So much of the work with previous data collections lacking these variables may have to be discounted.

The specifications of previous models have rarely taken explicit account of expected effect differences by family resource levels or tested for whether effects were constant or widened by grade. The latter consideration is critical to determining how models should be specified. For instance, if the conditions are present for a constant rather than an accelerating gap by grade, value-added models that control for previous years' test scores can show null effects of class size even though effects are present each year (Krueger 1999b). Effects would show up only in the first year in which class sizes were changed, but not in subsequent years. Such models would pick up only grade-by-grade acceleration in score changes. Here, simple cross-sectional models by grade without control for previous scores would show the total constant and cumulative effect to each grade.

The processes discussed above may or may not be the actual ones that exist in classes to produce class size effects. They simply point to the need to develop theories of the mechanisms underlying class size effects and to collect the data to test different theories. While a limited number of existing data sets might be able to start this process, it is difficult to see how definitive results are possible without more experimentation with more robust data collection on a much wider set of variables. Only by sorting this out can we be confident that models are specified correctly, estimation techniques are appropriate, and interpretations are accurate.

References

Advisory Group on the Scholastic Aptitude Test Score Decline. (1977). *On Further Examination.* New York: College Entrance Examination Board.

Achilles, C.M., Nye, B., Zaharias, J., and Fulton, B. (1993, January). *The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–91): A Legacy from Tennessee's Four-Year (K–3) Class Size Study (1985–90), Project STAR*. Paper presented at the North Carolina Association for Research in Education. Greensboro, North Carolina.

Armor, D. (1995). *Forced Justice: School Desegregation and the Law.* New York: Oxford University Press.

Atkinson, R., and Jackson, G. (Eds.). (1992). *Research and Education Reform: Roles for the Office of Educational Research and Improvement*. Committee on the Federal Role in Education Research, Washington, DC: National Academy Press.

Barnett, S. (1995, winter). Long-term Effects of Early Childhood Programs on Cognitive and School Outcomes, *The Future of Children 5*(3): 25–50.

Betts, J. R. (1997). *The Role of Homework in Improving School Quality,* Working manuscript. San Diego: University of California.

Betts, J. R., and Shkolnik, J. (1999a, summer). Estimated Effects of Class Size on Teacher Time Allocation in Middle and High School Math Classes, *Educational Evaluation and Policy Analysis 21*(2): 193–214.

Betts, J.R., and Shkolnik, J. (1999b). *The Effects of Class Size on Teacher Time Allocation and Student Achievement.* Working manuscript. San Diego: University of California.

Boruch, R. F. (1997). *Randomized Experiments for Planning and Evaluation.* Thousand Oaks, CA: Sage Publications.

Boruch, R. F., and Foley, E. (1998). The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials. In L. Bickman (Ed.), *Validity and Social Experimentation: Donald T. Cambell's Legacy.* Thousand Oaks, CA: Sage Publications.

Burtless, G. (1993). The Case for Social Experiments. In K. Jensen and P. K. Madsen (Eds.), *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*. Copenhagen: Ministry of Labour.

Burtless, G. (1996). *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington, DC: The Brookings Institution Press.

Cambell, J. R., Donahue, P. L., Reese, C. M., and Phillips, G. W. (1996). *NAEP 1994 Reading Report Card for the Nations and the States; Findings from the National Assessment of Educational Progress and Trial State Assessment* (NCES 95–045). Washington, DC: National Center for Education Statistics.

Cherlin, A. (1988). The Changing American Family and Public Policy. In A. Cherlin (Ed.), *The Changing American Family and Public Policy* (pp. 1–29). Washington, DC: The Urban Institute Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Consortium on Productivity in the Schools (1995). *Using What We Have to Get the Schools We Need*. New York: Teachers College Press, Columbia University.

Cook, M., and Evans, W. M. (1997). *Families or Schools? Explaining the Convergence in White and Black Academic Performance*. Working paper. University of Maryland.

Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 318–374). Washington, DC: The Brookings Institution Press.

Ferguson, R. F., and Ladd, H. F. (1996). How and Why Money Matters: An Analysis of Alabama Schools. In H. F. Ladd (Ed.), *Holding Schools Accountable*. Washington, DC: The Brookings Institution Press.

Finn, J. D., and Achilles, C. M. (1990, fall). Answers and Questions about Class Size: A Statewide Experiment. *American Educational Research Journal 27*(3): 557–577.

Finn, J. D., and Achilles, C. M. (1999, summer). Tennessee's Class Size Study: Findings, Implications and Misconceptions. *Educational Evaluation and Policy Analysis 21(*2).

Flynn, J. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological Bulletin 101*(2): 171–191.

Fuchs, V., and Rekliss, D. (1992). America's Children: Economic Perspectives and Policy Options. *Science 255:* 41–46.

Greenwald, R., Hedges, L. V., and Laine, R. D. (1996, fall). The Effect of School Resources on Student Achievement. *Review of Educational Research 66*(3): 361–396.

Grissmer, D. W. (1996). *Education Productivity*. Washington, DC: Council for Educational Development and Research.

Grissmer, D. W. (1999, summer). Assessing the Evidence on Class Size: Policy Implications and Future Research Agenda. *Educational Evaluation and Policy Analysis 21*(2): 231–248.

Grissmer, D. W. (forthcoming). The Use and Misuse of SAT Scores. *Journal of Psychology, Law and Public Policy.*

Grissmer, D. W., Flanagan, A., and Kawata, J. (1998). *Assessing and Improving the Family Characteristics Collected With the National Assessment of Educational Progress*. Santa Monica, CA: RAND.

Grissmer, D. W., Flanagan, A., Kawata, J., and Williamson, S. (forthcoming). *Improving Student Achievement: State Policies That Make a Difference*. Santa Monica, CA: RAND.

Grissmer, D. W., Flanagan, A., and Williamson, S. (1998a). Why Did the Black-White Test Score Gap Narrow in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.) *The Black-White Test Score Gap* (pp. 182–228). Washington, DC: The Brookings Institution Press.

Grissmer, D. W., Flanagan, A., and Williamson, S. (1998b). Does Money Matter for Minority and Disadvantaged Students?: Assessing the New Empirical Evidence. In W. Fowler (Ed.), *Developments in School Finance: 1997* (NCES 98–212) (pp. 13–30). U.S. Department of Education, Washington, DC: U.S. Government Printing Office.

Grissmer, D. W., Kirby, S. N., Berends, M., and Williamson, S. (1994). *Student Achievement and the Changing American Family*. Santa Monica, CA: RAND.

Grissmer, D. W., Kirby, S. N., Berends, M., and Williamson, S. (1998). Exploring the Rapid Rise in Black Achievement Scores in the United States (1970–1990). In U. Neisser (Ed.), *The Rising Curve: Long-Term Changes in IQ and Related Measures* (pp. 251–286). Washington, DC: American Psychological Association.

Hanushek, E. A. (1989). The Impact of Differential Expenditures on School Performance. *Educational Researcher 18*(4): 45–51.

Hanushek, E. A. (1994). *Making Schools Work: Improving Performance and Controlling Costs*. Washington, DC: The Brookings Institution Press.

Hanushek, E. A. (1996). School Resources and Student Performance. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: The Brookings Institution Press.

Hanushek, E. A. (1999, summer). Assessing the Empirical Evidence on Class Size Reductions from Tennessee and Nonexperimental Research. *Educational Evaluation and Policy Analysis, 21*(2): 143–163.

Hanushek, E. A., and Jorgenson, D. (1996). *Improving America's Schools: The Role of Incentives.* Washington, DC: National Academy Press.

Hanuskek, E. A., and Rivkin, S. G. (1997, winter). Understanding the Twentieth-Century Growth in U.S. School Spending. *Journal of Human Resources 32*(1): 35–67.

Hauser, R. M. (1998). Trends in Black-White Test Score Differentials: Uses and Misuses of NAEP/SAT Data. In U. Neisser (Ed.), *The Rising Curve: Long-term Changes in IQ and Related Measures* (pp. 219–250). Washington, DC: American Psychological Association.

Haveman, R., and Wolfe, B. (1994). *Succeeding Generations: On the Effects of Investments in Children*. New York: Russell Sage Foundation.

Haveman, R., and Wolfe, B. (1995, December). The Determinants of Children's Attainments: A Review of Methods and Findings. *Journal of Economic Literature 33*: 1829–1878.

Heckman, J. J., and Smith, J. (1995, spring). Assessing the Case for Social Experiments. *Journal of Economic Perspectives 9*(2): 85–110.

Heckman, J., Layne-Farrar, A., and Todd, P. (1996). Does Measured School Quality Really Matter? In H. F. Ladd (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (pp. 192–289). Washington, DC: The Brookings Institution Press.

Hedges, L. V., and Greenwald, R. (1996). Have Times Changed? The Relation between School Resources and Student Performance. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (pp. 74–92). Washington, DC: The Brookings Institution Press.

Hedges, L. V., Laine, R. D., and Greenwald, R. (1994). Does Money Matter? Meta-analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher 23*(3): 5–14.

Hedges, L. V., and Nowell, A. (1998). Black-White Test Score Convergence Since 1965. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 149–181). Washington, DC: The Brookings Institution Press.

Herrnstein, R. J., and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Jencks, C. (1992). *Rethinking Social Policy: Race, Poverty, and the Underclass*. Cambridge, MA: Harvard University Press.

Jencks, C., and Phillips, M. (Eds.) (1998). *The Black-White Test Score Gap.* Washington, DC: The Brookings Institution Press.

Jones, L. V. (1984, November). White-Black Achievement Differences: The Narrowing Gap. *American Psychologist 39:* 1207–1213.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., Sanders, M., and Chiesa, J. (1998). *Investing In Our Children: What We Know and Don't Know About the Costs and Benefits of Early Childhood Interventions.* Santa Monica, CA: RAND.

Karweit, N. (1989). Effective Kindergarten Programs and Practices for Students at Risk. In R. Slavin, N. Karweit, and N. Madden (Eds.), *Effective Programs for Students At Risk* (pp.103–142). Boston: Allyn and Bacon.

Koretz, D. (1986). *Trends in Educational Achievement.* Washington, DC: Congressional Budget Office.

Koretz, D. (1987). *Educational Achievement, Explanations and Implications of Recent Trend*s. Washington, DC: Congressional Budget Office.

Krueger, A. B. (1998, March). Reassessing the View That American Schools Are Broken. *Economic Policy Review*: 29–43.

Krueger, A. B. (1999a). An Economist's View of Class Size Reductions. Unpublished paper. Princeton University. Princeton, NJ.

Krueger, A. B. (1999b, May). Experimental Estimates of Education Productions Functions. *Quarterly Journal of Economics 114:* 497–532.

Ladd, H. F. (Ed.) (1996a). *Holding Schools Accountable.* Washington, DC: The Brookings Institution Press.

Ladd, H. F. (1996b). Introduction. In H. F. Ladd (Ed.), *Holding Schools Accountable* (pp. 1–22). Washington, DC: The Brookings Institution Press.

Lankford, H., and Wyckoff, J. (1996). The Allocation of Resources to Special Education and Regular Instruction. In H. F. Ladd (Ed.), *Holding Schools Accountable* (pp. 221–257). Washington, DC: The Brookings Institution Press.

Linn, R. L., and Dunbar, S. (1990). The Nation’s Report Card Goes Home: Good News and Bad News and Trends in Achievement, *Phi Delta Kappan 72*(2): 127–133.

Masten, A.S. (1994). Resilience in Individual Development: Successful Adaptation Despite Risk and Adversity. In M.C. Wang and E.W. Gordon (Eds.), *Educational Resilience in Inner City America: Challenges and Prospects*. Hillsdale: Lawrence Erlbaum Associates.

McLanahan, S., and Sandefur, G. (1994). *Growing Up with a Single Parent: What Hurts, What Helps*, Cambridge, MA: Harvard University Press.

Miller, K. E., Nelson, J. E., and Naifeh, M. (1995). *Cross-State Data Compendium for the NAEP 1994 Grade 4 Reading Assessment* (NCES 95–157). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., and Ehrle, K. (1999, summer). Evaluating a Pilot Program in Pupil/Teacher Reduction in Wisconsin. *Educational Evaluation and Policy Analysis 21*(2): 165–178.

Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades. *The Future of Children 5*(2): 113–127.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993). *NAEP 1992 Mathematics Report Card for the Nation and the States: Data from the National and Trial State Assessments* (NCES 93–261). Washington, DC: National Center for Education Statistics.

Murnane, R. J., and Levy, F. (1996). Evidence from Fifteen Schools in Austin, Texas. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: The Brookings Institution Press.

National Research Council (1989). *A Common Destiny: Blacks and American Society*. Washington, DC: National Academy Press.

Neisser, U. (Ed.). (1998). *The Rising Curve: Long-term Changes in IQ and Related Measures*. Washington, DC: American Psychological Association.

Nye, B., Hedges, L., and Konstantopoulos, S. (forthcoming). *The Effects of Small Class Size on Academic Achievement: The Results of the Tennessee Class Size Experiment.*

Nye, B., Hedges, L., and Konstantopoulos, S. (1999, summer). The Long-term Effects of Smaller Class Size: A Five-year Follow-up of the Tennessee Experiment. *Educational Evaluation and Policy Analysis 21*(2): 127–142.

Office of Science and Technology Policy. (1997, April). *Investing in Our Future: A National Research Initiative for America's Children for the 21st Century.* Washington, DC: The White House.

Orfield, G., and Eaton, S. (1996). *Dismantling Desegregation: The Quiet Reversal of Brown v. Board of Education*. New York: The New Press.

Phillips, M., Crouse, J. and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 229–272). Washington, DC: The Brookings Institution Press.

Popenoe, D. (1993). American Family Decline, 1960-1990: A Review and Appraisal. *Journal of Marriage and the Family 55*: 27–555.

Raudenbush, S. W. (1994). Random Effects Models. In H. Cooper and L. V. Hedges, (Eds.), *The Handbook of Research Synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W., and Wilms, J. D. (1995, winter). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics 20*(4): 307–335.

Reese, C. M., Miller, K. E., Mazzeo, J., and Dossey, J. A. (1997). *NAEP 1996 Mathematics Report Card for the Nation and the States* (NCES 97–488). Washington, DC: National Center for Education Statistics.

Rice, J. K. (1999, summer). Estimated Effects of Class Size on Teachers' Time Allocation in High School Mathematics and Science Courses. *Educational Evaluation and Policy Analysis 21*(2): 215–230.

Ritter, G., and Boruch, R. (1999, summer). The Political and Institutional Origins of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis 21*(2): 111–126.

Rock, D. A. (1987). The Score Decline from 1972–1980: What Went Wrong? *Youth and Society 18*(3): 239–254.

Rothstein, R. (1998). T*he Way We Were: The Myths and Realities of America's Student Achievement*. New York: The Century Foundation Press.

Rothstein, R., and Miles, K. H. (1995). *Where's the Money Gone? Changes in the Level and Composition of Education Spending*. Washington, DC: Economic Policy Institute.

Rutter, M. (Ed.). (1988). *The Power of Longitudinal Data.* Cambridge: Cambridge University Press.

Saranson, S. B. (1990). *The Predictable Failure of Educational Reform: Can We Change Course Before It's Too Late?* San Francisco, CA: Jossey-Bass Publishers.

Schofield, J. (1995). Review of Research on School Desegregation's Impact on Elementary and Secondary School Students. In J. A. Banks and C. A. McGee-Banks (Eds.), *Handbook of Research on Multicultural Education* (pp. 597–617). New York: McMillan Publishing.

Smith, M. S., and O'Day, J. (1991). Educational Equality: 1966 and Now. In D. Verstegen and J. Ward (Eds.), *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook* (pp. 53–100). New York: HarperCollins.

Stacey, J. (1993). Good Riddance to 'The Family': A Response to David Popenoe. *Journal of Marriage and Family 55*: 545–547.

United States Department of Agriculture (1997). *Estimates of the Costs of Raising Children*. Washington, DC: Author.

Wells, A., and Crain, R. (1994, winter). Perpetuation Theory and the Long-term Effects of School Desegregation. *Review of Educational Research 64*(4): 531–555.

Wilson, K. G., and Davis, B. (1994). *Redesigning Education*. New York: Henry Holt and Company.

Wilson, W. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago, IL: The University of Chicago Press.

Word, E., Johnston, J., and Bain, H. (1994). *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985–1990*. Nashville: Tennessee Department of Education.

Word, E., Johnston, J., and Bain, H. (1990). *Student Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study. Final Summary Report 1985–1990*. Nashville: Tennessee Department of Education.

Zill, N., and Rogers, C. (1988). Recent Trends in the Well-Being of Children in the United States and Their Implications for Public Policy. In A. Cherlin (Ed.), *The Changing American Family and Public Policy* (pp. 31–115). Washington, DC: The Urban Institute Press.

Response: Guidance for Future Directions in Improving the Use of NAEP Data

Sylvia T. Johnson[1]
Howard University

The issues of how to meaningfully use state National Assessment of Educational Progress (NAEP) data to assess and improve student achievement are the foci of the Grissmer and Flanagan and Raudenbush papers. They are exciting in their ramifications for future research and policy directions. The following discussion briefly describes NAEP and the whole idea of state-level data, then proceeds to review these papers in the context of their value in providing strategies for making these data more useful and informative assessments of national educational progress.

The assessment—as well as the improvement—of student achievement has long been a focus of educational policy at the state and the national levels and in the front lines of local school districts. With different emphases at different points in time, NAEP was originally designed as "the nation's report card" to provide information on student achievement in subjects widely taught in public schools, for the nation as a whole and for specific demographic and geographic subgroups. However, in the first iterations of NAEP back in the 1970s, the regional subgroups were large, each including several states. It was not until the introduction of the Trial State Assessment (TSA) in 1990 that a sampling and administration structure was developed which allowed for the direct comparison of states with one another. Such between-states comparisons were not the explicit intent of the program. Rather, the TSA was intended to allow each state to compare its performance with that of the nation as a whole or perhaps with similar states in its own geographic region.

[1] The author is Professor, Research Methodology and Statistics at the School of Education, Howard University, and a principal investigator for the Center for Research on the Education of Children Placed at Risk (CRESPAR), an OERI-funded research center. She may be reached at Howard University, 2900 Van Ness Street NW, 116 Holy Cross Hall, Washington, DC 20008.

Actually, the involvement of states in NAEP is not new. An association of state educators, along with a core working group of the nation's top psychometrics scholars, were involved in the original conception and planning of NAEP.[2] They wanted to implement a national assessment program to document student progress in a manner which would not pose a threat to lower-performing district participation. To help ensure a low-key, relatively nonintrusive assessment program, NAEP results reported the percent of students who correctly answered each "exercise," as the assessment items were termed. Results were reported for geographic areas, and national samples of students were identified at ages 9, 13, and 17. NAEP currently assesses samples of students in grades 4, 8, and 11, but also reports trends for both age and grade for cross-sectional samples from about 1970 to the present. The initial trend year varies according to the time at which trend samples were introduced in each subject matter area (Campbell, Voelkl, and Donahue 1997).

In fact, the actual implementation of the Trial State Assessment has had a marked effect on how we measure student achievement. First, a motivational effect seems apparent in the TSA scores: they are a bit higher than regular NAEP scores. Second, certain states were anxious about their comparative standings; therefore, the "multiple comparison charts" show which unadjusted state mean differences were statistically significant from one another, as well as which differences were in the range of what would be expected simply due to chance. In these tables, no adjustments were made for student and family characteristics or for school resources and teacher background differences, although the importance of these factors certainly had been demonstrated in research studies carried out by NAEP, as well as in analyses of NAEP data in the literature. The Raudenbush and the Grissmer and Flanagan papers both addressed this problem of more meaningful use of state TSA data to assess student achievement, and both papers focus on the Trial State Assessment of NAEP. The data are based on eighth grade mathematics proficiency estimates from the Trial State Assessment.

Response to the Raudenbush Paper

In his paper, Raudenbush proposed a synthesis of state results by developing models that include correlates of student proficiency within and between states. He began with a "Conceptual Model for State-level Policy Effects on

[2] Personal communication from W. E. Coffman, 1975.

Student Achievement" (figure 1), which shows state government acting on student achievement primarily through its effects on school resources and home backgrounds. His analysis thus began by using NAEP data to assess the contribution of school resources; e.g., school and teacher quality, and students' home background to student achievement. The analysis was done for each of the 40 participating states, thus providing within-state home and school correlates of mathematics proficiency for eighth graders. He found a substantial correlation between socially disadvantaged or ethnic minority status, parental education, and access to the key resources available for learning, specifically course-taking opportunities, positive school climate, qualified teachers, and cognitively stimulating classrooms. He noted that these findings are similar to other findings in the literature. Carrying this analysis to another level, Raudenbush found considerable variation in the patterns of these correlations across the 41 states participating in TSA. These findings provided estimates for the direct effects of schools and teachers on student achievement while controlling for the correlation between self-reported student background, school factors, and teacher practices.

Raudenbush's analytic approach offers far more useful information to states than the conventional means from the NAEP TSA. By comparing states on resources and educational opportunities, this work enables the examination of possible changes in policy that are likely to positively influence student achievement. This analysis utilizing hierarchical linear models demonstrates that only a small amount of residual variance exists between states that is not related to school resources and family background. It should be noted here that there is a wide range in the proportion of within-school variability within states, which Raudenbush points out is also apparent across states. But the within-state variation is worth the attention of individual states; and there is some work in this area, for example, William Cooley's paper on Pennsylvania (Beckford and Cooley 1993), which examines schools with sizable numbers of African American students in which these students score at or above the state mean on achievement measures, and which also cites other relevant investigations into these questions.

Given the demonstrated importance of school resources to achievement, the second part of the Raudenbush paper presented an examination of state differences in school resources. This work explores two questions: "Does the distribution of school resources likely reinforce or counteract inequalities arising from home environment? Do states differ, not only in the provision of resources, but also in the equity with which they are distributed?" This examination thus focused not just on resource differences between states, but also on how equitably resources are distributed within states. These within-state resource differences were then examined, not only in terms of student demographic background, but also in the interaction between resource differences, race-ethnicity, and parental education. Raudenbush's analysis and his illustrative plots (figures 4, 5, 6, and 7) show that the probability of access to key resources is a function of these background factors. Further, he found sizable differences attributable to student background, education, and ethnicity; and these differences were associated with access to resources related to school success. For example, factors such as teacher quality and experience, school climate, whether or not a school offered algebra, and whether students were assigned to math teachers who majored in math and who emphasized reasoning in their classroom instruction—these were related to success, as well as to the variables used as predictors in Phase I of the Raudenbush work. These same variables, when examined across states, result in the ellipses (figures 8 and 9) that visually show the relative access to resources provided to African American students. For example, figure 8 shows that in South Carolina and Mississippi, having parents who are college educated offers only modest advantage to students, and the advantage is about the same for African Americans and youth of other backgrounds.

The meaning and the utility of Raudenbush's findings and this methodology for states and districts, and perhaps also for schools, are substantial. First, the absence of large statistical differences between adjusted means should in no way encourage states that they are to do little. The kind of action that would be prompted from the states was a concern of the National Assessment Governing Board when it decided, when plans were made for reporting the 1990 results, to report unadjusted means only. Second, the Raudenbush analyses clearly show the importance of access to school resources for student achievement. In order to move toward equity for all students, these findings demonstrate unequivocally that resource accessibility is a key factor.

In terms of the current "affirmative action" debate, these findings also have important implications for many states, such as Florida and California. First, can a state logically expect proportional representation by ethnicity on college entry characteristics such as test scores and course-taking patterns when it has systematically limited access to resources available to students? Given the demonstrated relation of resources to measured achievement reported in the Raudenbush paper, it would seem only logical that a state, in making college entrance decisions, should take into consideration the relative access it provides to certain resources in elementary and secondary school. Such a procedure would need, certainly, a well-specified model and additional research.

Though the many issues are still in flux, reactions to affirmative action are often far too simplistic in their conceptions of how the specific mechanisms work in practice. Where race has served as a factor in the allocation of skilled teachers and other education resources, whether by design or not, this condition raises the question of how this method of allocation has to be considered when the measurable results of such allocations are evaluated.

The author's broad recommendations should be noted here. If student progress and subject matter proficiency are examined on a year-to-year basis, the extent to which the educational system (or even the school) differentially provides resources that support student progress should also be examined. This is the *opportunity-to-learn* concept. The soundly based but creative methodology that is employed in the Raudenbush work offers strong promise for helping us better understand how student background and resource access are interrelated, as well as when such access factors have been modified. The latter could be examined by extending the analysis over time so that the progress of states in modifying resource access could be followed and appraised. The author suggested extending the collection of data by NAEP to measures of resources and development indicators from these data. Such a direction seems logical and important, but it stands in opposition to current plans to release results more quickly and to collect fewer data from students, teachers, and schools than has been done in the past.

How can states and schools use these findings? To begin with, they can collect comparable data at the district and school levels so that the internal allocation of resources is more completely documented. They can also modify teacher assignments to provide more equitable distribution of highly skilled teachers, although such a goal may entail the need for financial incentives, along with improving the facilities and working conditions for teachers in some schools, and other strategies. For teachers, the range in access to *everything*—from professional development to clean bathrooms—is very great across schools, even in the same or nearby districts. State officials may lobby for increased state support to remedy access problems: they could target selected schools and districts, using these findings as a basis for the request. They may investigate how parental education operates to increase access to resources and provide help to parent groups in lower-access schools to develop strategies to bring about change for the better in their schools. They can recognize the broad scope of the documented inequities and develop a broad-based strategy, sustained over the long term, to simultaneously and progressively change the inequities in resource allocation that are so widely spread among states, systems, and schools.

Response to the Grissmer and Flanagan Paper

In their paper, "Moving Educational Research toward Scientific Consensus," Grissmer and Flanagan assert the need to improve the consistency and accuracy of results in educational research so that a basic knowledge base in education can be built, one that can be accepted by a diverse research community as well as by educational practitioners and policymakers. The authors assert that improving nonexperimental data, along with the associated methodology, may not be enough to achieve consensus. Rather, they contend that experiments are needed and, further, that experiments should often employ models such that the size of a given year's effect can be viewed as dependent on the current year's and previous years' effects. These findings should then be used to build micro-theories of educational process.

Grissmer and Flanagan dealt with the issue of the relation between school resources and school achievement in two major examples. The first example which they cite is the rapid change in NAEP scores of black students, especially from 1970 to 1988 or 1990. In the case of black student progress in NAEP scores, Grissmer and Flanagan gave credit to compensatory and developmental programs and school desegregation activities, but they did not offer conjecture regarding the score declines that occurred starting from 1988 or1990.

The Tennessee class size study is well presented by Grissmer and Flanagan; and the posing of a simple process model for these effects is a useful and important addition to the literature. The detail presented to amplify and explain how teacher reactions to the scarcity or abundance of class time may interact with student characteristics is thoughtful. Readers are encouraged to take the time to examine that part of the paper carefully, as it provides further support for the need of extensive data collection, either from teacher questionnaires, interviews, or classroom observations.

It is good to see the Tennessee class size experiment get the kind of attention toward implementation that it has deserved for some years. This long-term experiment received relatively little attention in the policy arena until recently, but now seems to be getting wide notice. Certainly, the Tennessee class size study has reached the ear of the President—it is, indeed, a factor in the call for many new teachers across the nation. Interestingly, the Tennessee study results from collaboration between the historically black university, Tennessee State University, and the State Department of Education, with important guidance from a participant in this conference, Jeremy Finn. This well-designed study made it possible to study the effects of smaller classes over time. Thus, it is an excellent model for the kind of work that needs to be done to develop and test theories in education.

In addition to their suggestions for extending and improving NAEP data collections, Grissmer and Flanagan suggest improving NAEP by collecting family characteristics at the school level, possibly using Census Bureau data to augment NAEP. There are some states and other jurisdictions which have used Census data and other federal reporting information to improve their estimates for allocation of social and economic services. A major problem in this work has been the adequacy of geo-coding (the coding of addresses and other location information) in the files proposed for use to improve estimation. The problem is especially severe in sparsely populated areas; namely, rural communities, older urban industrial zones that have lost population with the closing of plants, and small towns. A National Research Council panel has been examining problems of estimation of poverty in small geographic areas, and the U.S. Department of Education, through the National Center for Education Statistics, is a sponsor of this work. Working in close collaboration with the Census Bureau, the panel has been operating for about 3 years, has published three interim reports, and is working to complete a final report (National Research Council 1999). Their findings should be useful in the improvement of parameter estimation for many forms of resource allocation.

Grissmer and Flanagan also suggested a school district sample rather than a school sample for NAEP and the use of a longitudinal cohort. A district sample, though more expensive to collect, might enable comparisons of scores for urban and suburban districts within metropolitan areas of similar size, though this level of reporting is disallowed under current NAEP authorization. Now, of course, prior to TSA, NAEP was a low-stakes testing program. Since comparisons are now easily made between states, it is worth considering whether we might facilitate between-district comparisons, if solid methodology that gives consideration to resource allocation is used to interpret those differences. Such a change would raise the stakes for State NAEP as well as for national NAEP. Given the effect of other high-stakes testing programs and the fundamental role of NAEP as the nation's report card, such changes should be carefully reviewed.

The authors support the use of longitudinal cohorts. A longitudinal study would involve identifying students or at least forming blocks at the school or district level, but the advantages would need to be weighed along with costs. NAEP currently examines trends by retaining a common core of test items which are administered to cross-sectional grade level groups.

Conclusion

Now let us consider what these papers, taken together, tell us. Both studies point out the importance of a school's climate and culture to discipline. In our work at CRESPAR, the Center for Research on the Education of Students Placed at Risk, an OERI-funded research center located at Howard University and Johns Hopkins University, we are guided by a talent development model recently articulated in an article by my colleague, Serge Madhere (1998; see also Boykin 1996). This model has a number of points, the most important one of which is that all children can learn, given adequate opportunity and that their backgrounds and culture have strengths that can be built on to motivate and encourage student learning. Learning is more a function of coherent instruction than of a child's social origin. Motivation begets greater learning, which begets greater motivation. Nurturing is the key to motivation, especially at difficult transition points.

Both of these papers offer creative methodological approaches to the use of state-level data for school improvement and for theory building. Both demonstrate the importance of resource allocation for student achievement and the interaction between race and resources, and both imply procedures for increasing proficiency among African American students. Both imply the need for more complex NAEP data at the level of the student, the family, the teacher, the school, and the state. This emphasis, however, runs counter to the current push to simplify and speed up the data collection and reporting process. More complex data require more complex consideration before developing conclusions.

Both studies show the need to better understand the why's and how's of improving student achievement. For example, how do teachers use time? When they have more time per student, what are the features of resources that make them effective in influencing achievement?

These are important directions for researchers and policymakers to consider. Fortunately, there is much in these papers to help guide that progress.

References

Beckford, I. A., and Cooley, W. W. (1993). *The Racial Achievement Gap in Pennsylvania*. Pennsylvania Educational Policy Studies Paper Number 18. ERIC No. EDD 389760.

Boykin, A. W. (1996, April). *A Talent Development Approach to School Reform: An Introduction to CRESPAR*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Campbell, J. R., Voelkl, K. E., and Donahue, P. L. (1997). *NAEP 1996 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969 to 1996, Mathematics, 1973 to 1996, Reading, 1971 to 1996, Writing, 1984 to 1996*. (NCES 97–985). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Madhere, S. (1999). Cultural Diversity, Pedagogy, and Assessment Strategies. *Journal of Negro Education 67*(3): 280–295.

National Research Council. (1999). *Estimation of Poverty in Small Geographic Areas* (Third Interim Report). Washington, DC: National Academy of Science Press.

Section II. Using Longitudinal Data to Assess Student Achievement

Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data 103
Meredith Phillips

Certification Test Scores, Teacher Quality, and Student Achievement 133
Ronald F. Ferguson with Jordana Brown

Response: Two Studies of Academic Achievement 157
Robert M. Hauser

Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data

Meredith Phillips
University of California, Los Angeles

In 1966, James Coleman published results from the first national study to describe ethnic differences in academic achievement among children of various ages. Since that time, we have made considerable progress in survey design, cognitive assessment, and data analysis. Yet we have not made much progress in understanding when ethnic differences in academic achievement arise, how these differences change with age, or why such changes occur.[1] The purpose of this paper is to highlight several reasons why we have learned so little about these important issues over the past few decades. I begin by reviewing recent research on how the test score gap between African Americans and European Americans changes as children age. I then discuss several conceptual and methodological issues that have hindered our understanding of ethnic differences in academic achievement. I raise these issues in the hope that we will make more progress toward eliminating the test score gap during the next decade than we have during the last.[2]

[1] I use the term "ethnic" to refer to the major ethnic and racial groups in the United States (namely, African Americans, European Americans, Latinos, Asian Americans, and Native Americans). Whenever the samples are large enough, I also consider variation within these socially constructed categories (for example, differences between Mexican Americans and Puerto Rican Americans).

[2] I thank Robert Hauser, Larry Hedges, Christopher Jencks, Jeff Owings, and Michael Ross for their comments on an earlier draft. I did not make all the changes they suggested, however; and they are in no way responsible for my conclusions. Please direct all correspondence to Meredith Phillips, School of Public Policy and Social Research, UCLA, 3250 Public Policy Building, Los Angeles, CA 90095–1656 or phillips@sppsr.ucla.edu.

Does the Achievement Gap Change as Children Age?

My colleagues and I recently analyzed data from a number of national surveys in order to estimate how the achievement gap changes as children age (see Phillips, Crouse, and Ralph 1998). Answering this question can help us understand the potential causes of the gap. Suppose, for example, that the black-white gap did not widen at all after first grade, even among black and white children who began school with similar skills. If that were the case, we might conclude that families, communities, preschools, or kindergartens were mainly responsible for the gap. On the other hand, suppose that the black-white gap did widen between the first and the twelfth grades, even among children who started school with similar scores. If that were the case, we might conclude that schools were mainly responsible for the gap. As it turns out, the "truth" seems to fall somewhere between these extremes.

Cross-sectional Results

One way to describe age-related changes in the black-white gap is to estimate the size of the gap in as many surveys as possible and then combine these estimates. We have done this with the national surveys listed in table 1. Figure 1a arrays the black-white math gaps from these surveys by age. The lines around the estimates show their precision. We can also array these gaps by year of birth, which shows the historical trend in the black-white math gap (see figure 1b). Because the black-white gap narrowed during the 1970s and 1980s, however, we need to make sure that age-related changes in the gap are not confounded with historical changes. In order to disentangle the effects of age from the effects of history, we estimated a multivariate model that controlled for the historical trend while estimating the age-related trend.[3] Table 2 presents these results. It shows the following: the black-white math gap widens by about 0.18 standard deviations between the first and the twelfth grades; the reading gap stays relatively constant; the vocabulary gap widens by about 0.23 standard deviations.[4] A gap of one standard deviation on the math or verbal SAT is 100 points. Therefore, our cross-sectional results imply that the black-white math and vocabulary gaps widen by the equivalent of just under 2

[3] For details on the sample and analysis, see Phillips, Crouse, and Ralph (1998).

[4] To obtain these estimates, multiply the coefficients in the first row of table 2 by 12 years of school.

Table 1. Data Sets Used in Meta-analysis

Acronym	Name	Test Year(s)	Grades Tested
EEO	Equality of Educational Opportunity Study	1965	1,3,6,9,12
NLSY	National Longitudinal Survey of Youth	1980	10,11,12
HS&B	High School & Beyond	1980	10,12
LSAY	Longitudinal Study of American Youth	1987	7,10
CNLSY	Children of the National Longitudinal Survey of Youth	1992	Preschool, K, 1,2,3,4,5
NELS	National Education Longitudinal Study	1988, 1990, 1992	8,10,12
PROSPECTS	Prospects: The Congressionally-Mandated Study of Educational Growth and Opportunity	1991	1,3,7
NAEP	National Assessment of Educational Progress	1971–1996	4,8,11

Figure 1a.
Standardized Black-White Math Gaps, by Grade Level

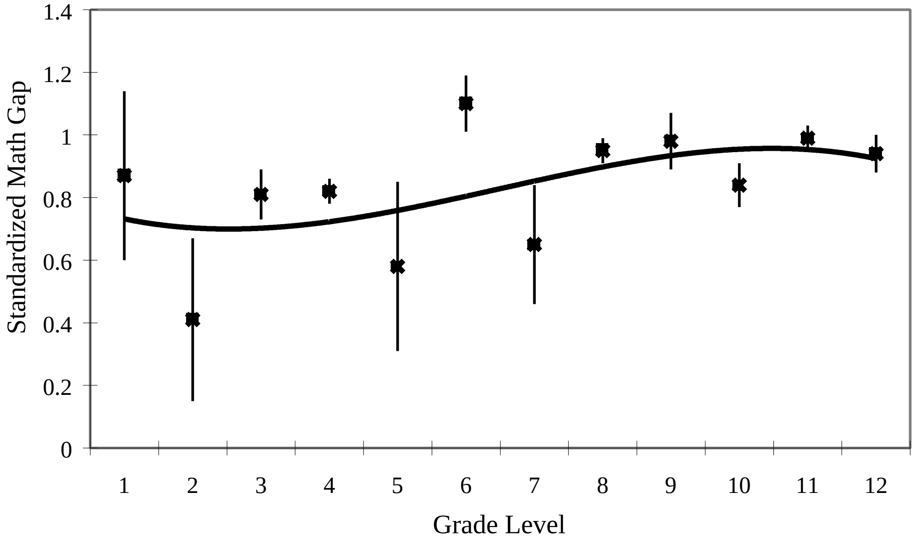


Figure 1b.
Standardized Black-White Math Gaps, by Year of Birth

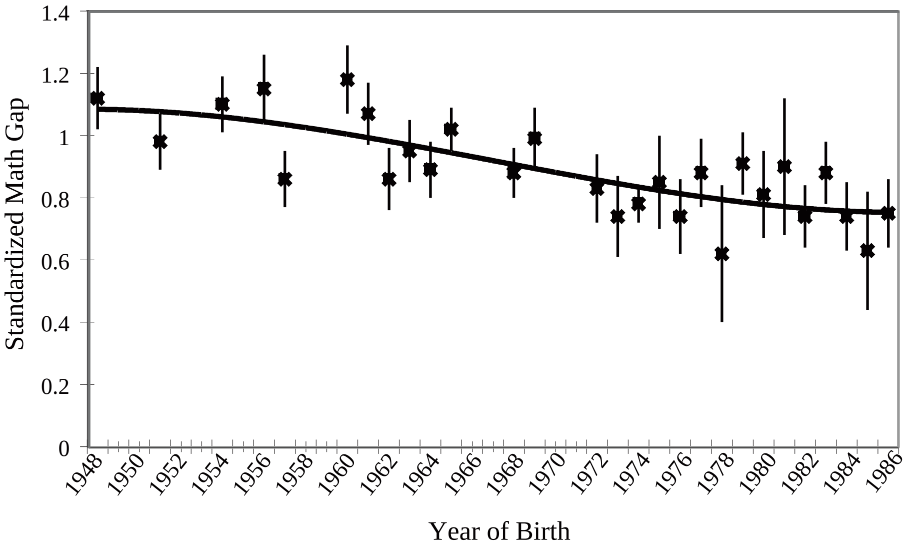


Table 2. Effects of Grade at Testing and Year of Birth on Black-White Test Score Gaps

		Dependent Variables					
		Mathematics (*N*=45)		Reading (*N*=45)		Vocabulary (*N*=20)	
Independent Variables		1	2	1	2	1	2
Grade level	B	.015	...	.002	...	.019	...
	SE	(.004)		(.006)		(.006)	
Grades 1–6	B	...	.051	...	-.011	...	.034
	SE		(.014)		(.023)		(.012)
Grades 7–8	B	...	-.054*	...	.016	...	.025
	SE		(.028)		(.051)		(.032)
Grades 9–12	B	...	.021*	...	.010	...	-.018
	SE		(.013)		(.024)		(.017)
Month of testing	B	-.011	-.007	.003	.000	.015	.011
	SE	(.004)	(.004)	(.005)	(.007)	(.018)	(.018)
Year of birth before 1978	B	-.014	-.014	-.020	-.020	-.010	-.011
	SE	(.002)	(.002)	(.002)	(.002)	(.003)	(.003)
Year of birth after 1978	B	.002*	.004*	.020*	.018*	.031*	.039*
	SE	(.006)	(.005)	(.009)	(.010)	(.011)	(.012)

Table 2. Effects of Grade at Testing and Year of Birth on Black-White Test Score Gaps (continued)

		Dependent Variables					
		Mathematics (*N*=45)		Reading (*N*=45)		Vocabulary (*N*=20)	
Independent Variables		1	2	1	2	1	2
Longitudinal survey	B	-.039	-.043	-.069	-.063	-.346	-.273
	SE	(.033)	(.033)	(.047)	(.051)	(.157)	(.161)
IRT metric	B	.175	.149	.159	.174	.068	.000
	SE	(.033)	(.035)	(.046)	(.051)	(.082)	(.088)
Intercept	B	.765	.653	.746	.792	.889	.833
	SE	(.034)	(.054)	(.056)	(.092)	(.049)	(.057)
Adjusted R^2		.790	.815	.693	.680	.745	.806

NOTE: The dependent variables are standardized black-white gaps (i.e., $(\overline{W}-\overline{B})/SD_T$) computed from the surveys listed in table 1. The actual data appear in table 7A–1 in Phillips, Crouse, and Ralph (1998). Standard errors are in parentheses. The spline coefficients for grade level and year of birth show the actual slope for that spline. The spline standard error indicates whether the slope differs from zero. * indicates that the spline's slope differs significantly from a linear slope at the .05 level. Each gap is weighted by the inverse of its estimated sampling variance. See Phillips, Crouse, and Ralph (1998) for details on the other variables in this analysis. See pp. 118-19 of Pindyck and Rubinfeld (1991) for an introduction to spline (piecewise linear) models. See Cooper and Hedges (1994) for details on the meta-analytic methods used in this analysis.

SAT points a year, or by 18 to 23 SAT points over the course of elementary, middle, and high school.

These cross-sectional estimates have two advantages over longitudinal estimates. First, the data span nearly all grade levels, from early elementary school through late high school. No national longitudinal survey has ever tested children over an interval spanning both elementary school and high school. Second, because cross-sectional surveys do not follow students over time, they are less subject to attrition and thus tend to be more nationally representative than longitudinal surveys. A problem with our cross-sectional results, however, is that they combine data on children from different samples, who were assessed on different, possibly incomparable, tests. Another problem is that cross-sectional data cannot tell us whether the black-white gap widens among children who start school with the same skills. That question, which is central to the concern that schools may not be offering black and white students equal educational opportunities, can be answered only with longitudinal data.

Longitudinal Results

During the late 1980s and early 1990s, two national longitudinal surveys assessed students multiple times as they moved through school. The National Education Longitudinal Survey (NELS) is the more familiar of these studies. NELS is a large national survey that first tested eighth graders in 1988 and then retested them in 1990 and 1992. Prospects, a survey of two cohorts of elementary school students and one cohort of middle school students that began in 1991, is less familiar than NELS because it is not yet readily available to researchers. The Prospects data were collected mainly to evaluate the effectiveness of Chapter 1 (now Title I), but their secondary purpose was to describe yearly achievement growth during elementary and middle school. The youngest of the Prospects cohorts was first tested at the beginning of first grade and followed through the end of third grade. The middle Prospects cohort was first tested at the end of the third grade and followed through the end of sixth grade. The oldest cohort was tested at the end of seventh grade and followed through the end of ninth grade.

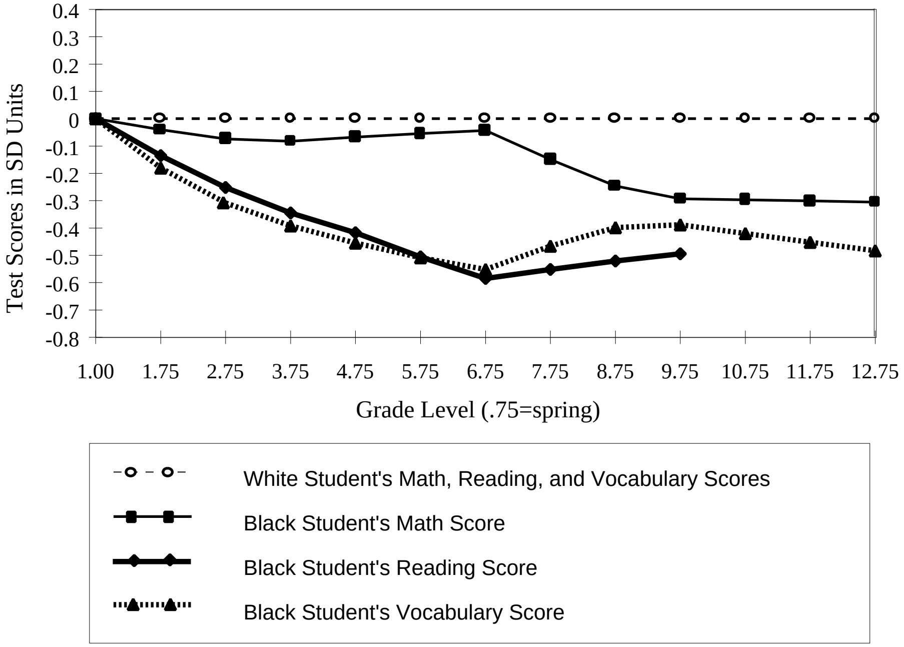
In order to understand achievement growth over an interval longer than four years, we have to piece together data from these different cohorts. My colleagues and I have used these data to estimate whether black children who start out with the same skills as whites learn less over the school years.[5] Our estimates are very imprecise because the Prospects sampling design was relatively inefficient and because we do not have data for every school year.[6] Nonetheless, our results suggest that African American children fall somewhat behind equally skilled white children, particularly in reading comprehension, and particularly during the elementary school years (see figure 2).[7] Taken together, we estimate that at least half of the black-white gap that exists at the end of twelfth grade can

[5] See Phillips, Crouse, and Ralph (1998) for details.

[6] Also, a very large percentage of the Prospects students left the study before the second and third waves. When Phillips, Crouse, and Ralph (1998) compared cross-sectional and longitudinal samples drawn from Prospects, however, they found that the mean black-white gap differed by less than 0.05 standard deviations across all tests. And although the longitudinal samples were more advantaged than the cross-sectional samples, *racial* differences in attrition were small and mostly involved regions of residence and urbanism. See chapter 3 of Phillips (1998) for more on nonrandom attrition in both Prospects and NELS.

[7] See Phillips, Crouse, and Ralph (1998) for a comparable figure using cross-sectional data, as well as for a figure that shows the imprecision of these predictions.

Figure 2. Predicted Test Scores for Two Students, One Black, One White, Who Both Started First Grade with True Math, Reading, and Vocabulary Scores at the Mean of the Population Distribution



be attributed to the gap that already existed at the beginning of first grade. The remainder of the gap seems to emerge during the school years.

This widening of the gap may not be attributable to schooling *per se*, however. Because of summer vacation, students spend only 180 days a year in school. Because neither Prospects nor NELS tested children in the fall and the spring of each school year, it is impossible to know how much of the gap that emerges over the course of schooling should be attributed to schools and how much should be attributed to summer vacations.[8]

In an ideal world, we would know precisely when ethnic differences in test scores first emerge and how they develop during the preschool years.

[8] Several other studies have examined summer learning patterns (e.g., Cooper et al. 1996; Entwisle and Alexander 1992, 1994; and Heyns 1978, 1987). Further, Prospects tested an unrepresentative subsample of students in the fall and spring of first and second grade. I review these results later in the paper.

We would also know how ethnic differences change both every school year and every summer. This information would help us identify the most important reasons why African American and Latino children score lower than whites and Asian Americans on math and reading achievement tests. Unfortunately, we are not close to knowing the answers to these seemingly basic, descriptive questions. In the remainder of this paper, I discuss several explanations for this knowledge gap.

Why Do We Know So Little?

The most obvious reason why we have made so little progress on the test score gap puzzle since 1966 is that most researchers have been reluctant to study it. Rather than directly tackling this politically sensitive subject, most scholars have tried to understand ethnic inequalities in academic skills by comparing socioeconomically disadvantaged students to advantaged students, by comparing students in high poverty schools to those in low poverty schools, or by comparing urban students to suburban students. All these comparisons pose interesting questions for social science. None, however, brings us closer to understanding ethnic differences in academic achievement because ethnicity does not overlap with social class and urbanism as much as most researchers assume.

Table 3 illustrates this problem. It shows the magnitude of the black-white test score gap among a national sample of eighth graders, according to the education and income levels of their parents, as well as the poverty and urbanism of their schools. If these other variables were adequate substitutes for race, the black-white gap would disappear after these variables were taken into account. The black-white gap does shrink, but it is still large *within* each of these categories. More sophisticated analyses that simultaneously control many family background variables yield similar results (see Phillips et al. 1998).[9] Racial and ethnic differences in test scores are *not* the same as SES differ-

[9] See also appendix C of Phillips, Crouse, and Ralph (1998) for data on how much the black-white gaps in Prospects and NELS shrink after controlling a number of common indicators of family background.

ences. Therefore, if we want to understand and eliminate racial and ethnic differences in test scores, we need to confront the problem directly.[10]

The results I presented on age-related changes in the black-white test score gap illustrate that even when we decide to focus explicitly on ethnic differences in academic skills, however, the available data are inadequate. In order to improve our understanding of the development and causes of ethnic differences in academic skills, supporters of educational surveys, such as the National Center for Education Statistics (NCES), need to do the following:

- Focus primarily on preschool and elementary school students rather than high school students;
- Assess at least a subsample of students in the fall and spring of *every* school year;
- Maximize measurement variation *within* each survey;
- Fund more than one survey of the same population at a time; and
- Remember that, because education begins before formal schooling, education surveys must also do the same.

I will elaborate on each of these points in turn.

The Importance of Early Schooling

Scholars who have studied ethnic differences in test scores have mostly focused on adolescents.[11] This is a mistake, for reasons that I will illustrate. It is, however, a reasonable mistake—at least among quantitative scholars—be-

[10] This does not mean, of course, that policies aimed at reducing the test score gap need be targeted at specific racial or ethnic groups rather than at low-scoring students in general. As Christopher Jencks and I argue in our introduction to *The Black-White Test Score Gap* (1998), the best policies seem to be those that help both blacks and whites, but help blacks more.

[11] See, for example, Ogbu (1978), Fordham and Ogbu (1986), Kao, Tienda, and Schneider (1996), Cook and Ludwig (1997), and Ainsworth-Darnell and Downey (1998). The main exceptions are Doris Entwisle and Karl Alexander, the founders of the Beginning School Study (BSS) in Baltimore, who continue to follow students who began first grade in 1982. Entwisle and Alexander's (1988, 1990, 1992, 1994) studies have been the source of most of our knowledge about the development and causes of black-white differences in academic achievement. Yet the BSS sample is relatively small, does not include enough schools to estimate between-school differences precisely, does not include Latinos or Asians, and may not generalize to other samples, because white students in the Baltimore public schools are not representative of white students nationally.

Table 3. Standardized Black-White Test Score Gap among Eighth Graders in NELS, by Parental Education, Income, School Poverty, and Urbanism

	Black-white gap on a combined math and reading test among eighth graders
In the overall population:	-0.80
Whose parents are:	
High school dropouts or graduates	-0.55
College graduates	-0.85
Whose parents' income is in the:	
Bottom fifth of the distribution	-0.53
Top fifth of the distribution	-0.53
Who attend schools in which:	
More than 40 percent of students are eligible for free or reduced lunch	-0.57
Fewer than 5 percent of students are eligible for free or reduced lunch	-0.65
Who attend:	
Urban schools	0.89
Suburban schools	-0.75

NOTE: The denominator of the gap is the weighted overall population standard deviation.

cause NCES has not, until very recently, supported surveys of elementary school students. NCES has conducted three large surveys of high school students: The National Longitudinal Study of 1972 (NLS-72), the High School and Beyond survey of 1980 (HSB), and NELS. All the data and documentation from these three studies are available to researchers who agree to abide by specific security provisions. Although the U.S. Department of Education has also supported two longitudinal surveys of elementary school students—the Sustaining Effects Study of 1976 and the Prospects study—the main purpose of these elementary school surveys was to evaluate Chapter 1, not to study learning during elementary school. Neither data set is widely available to the research community, and the quality of the data and documentation is much lower than in the high school surveys.[12]

[12] The main benefit of having conducted a new high school survey every 10 years is that the designers of each new survey are able to learn from mistakes made in the previous survey. The HSB improved on the NLS:72, and NELS improved considerably on the HSB. Prospects did not benefit much from the mistakes made in collecting the Sustaining Effects data because the Sustaining Effects data were never widely available enough to be subjected to close scrutiny. (The designers of Prospects did benefit, however, from improvements in the high school surveys—NELS, in particular.)

Lower Year-to-year Correlations

Two empirical facts illustrate why we need to focus on early schooling if we want to understand ethnic differences in achievement. First, the correlations between students' scores in one year and their scores in the following year tend to be lower during elementary school than during high school, even after correcting for measurement error (see table 4). In fact, year-to-year correlations are so high during high school (0.98 for both reading and math after correcting for measurement error) that hardly any students change their relative rank between the eighth and the twelfth grades. The lower year-to-year correlations during elementary school imply that cognitive skills are most malleable among young children and suggest that interventions aimed at changing students' skills may be most successful in the early school years.

Table 4. Year-to-year Correlations in Prospects and NELS

	Observed Correlation	True Correlation
Prospects		
Grade 1, Fall to Spring		
Reading	.49	.64
Math	.67	.81
Grade 3, Spring to Spring		
Reading	.69	.75
Math	.71	.80
NELS		
Grade 8, Spring to Spring		
Reading	.90	.98
Math	.94	.98

NOTE: Correlations are based on weighted data. They are disattenuated using the reliabilities reported in the Prospects Interim Report (1993) and the Psychometric Report for the NELS: 88 Base Year Through Second Follow-Up.

Lower Gain-Initial Score Correlations

A second, and related, reason to study achievement during elementary school is that the correlation between children's initial skills and how much they learn as they age is lower in elementary school than in high school. Everyday empiricism often suggests that additional advantages typically befall those who are already most advantaged. In terms of cognitive skills, this "fan spread" theory implies that students who have the best reading and math skills when they begin school will tend to gain the most reading and math skills as they move through school. Conversely, students who begin school with the fewest reading and math skills will tend to gain the fewest skills as they move through school. As students' skills diverge, their growth trajectories will come to resemble an opening fan.[13]

Scholars have debated for decades whether fan spread applies to test scores. For fan spread to occur, students' test score gains have to be positively correlated with their initial scores. More than 30 years ago, Benjamin Bloom (1964) argued in his famous study, *Stability and Change in Human Characteristics*, that gains on a wide variety of tests were uncorrelated with initial scores. But many scholars dismissed his findings as a result of measurement error in the tests (see, for example, Werts and Hilton 1977).[14] David Rogosa and John Willett (1985) have shown, however, that the relationship between gains and initial scores depends both on the shape of individual growth curves for a particular skill and on the age at which researchers measure initial status on that skill. The "true" correlation between initial status and gains can therefore be negative, positive, or 0, depending on what the particular test measures and when the children taking the test normally learn the skills that the test measures.

Table 5 shows that, even after correcting for measurement error, the correlations between gains and initial scores tend to be lower among elementary school students than high school students. The true gain-initial score correlations of around 0.50 during the first two years of high school indicate that the

[13] The fan-spread phenomenon is also known as the Matthew effect, after the biblical "For he that hath, to him shall be given… but he that hath not, from him shall be taken away…."

[14] Measurement error often creates a negative correlation between gains and initial scores. This is because, if we measure gains by subtracting observed Time 1 scores from observed Time 2 scores, the random error term in the Time 1 score appears with a positive sign in the Time 1 score, but with a negative sign in the gain score.

Table 5. Correlations between Initial Scores and Gains in Prospects and NELS

	Observed Gain, Initial Score Correlation	True Gain, Initial Score Correlation
Prospects		
Fall of Grade 1 to Spring of Grade 2		
Reading	-.33	-.14
Math	-.43	-.30
Spring of Grade 3 to Spring of Grade 5		
Reading	-.49	-.45
Math	-.43	-.36
NELS		
Spring of Grade 8 to Spring of Grade 10		
Reading	-.09	.53
Math	.06	.48
Spring of Grade 10 to Spring of Grade 12		
Reading	-.28	-.12
Math	-.08	.23

NOTE: Correlations are based on weighted data. They are disattenuated using the reliabilities reported in the Prospects Interim Report (1993) and the Psychometric Report for the NELS:88 Base Year Through Second Follow-Up and equations 11 and 13 in Willett (1988).

highest-scoring students at the end of eighth grade tend to gain the most during ninth and tenth grade. This is, of course, exactly what we would expect if students were tracked based on their math and reading scores and if students in higher-level math and English courses learned more math and reading skills. But these results contrast markedly with those for elementary school.[15] The correlations between true gains and true initial scores in Prospects tend to be negative. This implies that students who start first grade with higher scores tend to learn less between first and third grade than those who start first grade with lower scores.

[15] Note, however, that correlations between gains and initial scores are lower during the last two years of high school, especially for reading. Jencks and Phillips (1999) found near zero correlations between gains and initial scores in the High School and Beyond (HS&B) during the last two years of high school. If the NELS and HS&B tests measured all the skills that high-scoring students learn in the last two years of high school (e.g., calculus), the gain-initial score correlations might be as large and positive as they are for the first two years of NELS.

The negative correlation between gains and initial scores in Prospects suggests that elementary schools may be structured (intentionally or not) to help those children who have the weakest academic skills.[16] For example, first grade teachers may focus more on teaching children how to read than on improving the reading skills of those who already know how to read. Or perhaps elementary school math textbooks are geared to below average students at each grade level, so that students with the weakest math skills are challenged the most. Of course, reforms aimed at enriching the elementary school curriculum may end up improving learning among the most highly skilled students, thereby increasing the gap between high and low scorers.

Prospects is not the only data set to suggest that elementary schools may reduce rather than exacerbate the test score gap. Entwisle and Alexander have reported similar results for the BSS (for a recent example, see Alexander and Entwisle 1998). Because the black-white test score gap seems to widen more during elementary school than during high school, because children's scores are less fixed during elementary school than during high school, and because elementary schools may already help reduce the gap between high and low scorers, studies involving elementary school-age children are essential for understanding the development of ethnic differences in achievement. Fortunately, NCES is funding the Early Childhood Longitudinal Study (ECLS-K), which began collecting data on a national sample of kindergartners in the fall of 1998. The ECLS-K is much needed. One aspect of its design, however, will severely limit its contribution to our knowledge about ethnic differences in academic achievement.

Measuring Summer Learning

The ECLS-K currently plans to test a 25 percent subsample of students in the fall of first grade. For these children, the ECLS-K will be able to distinguish learning during kindergarten and first grade from learning during the summer

[16] An alternative explanation for the low gain-initial score correlation in elementary school is that the elementary school tests do not measure the skills that high-scoring students acquire during elementary school. This is almost certainly true. Yet, it must be more true for tests administered during elementary school than for tests administered during high school in order for it to explain why the gain-initial score correlations are so much lower in elementary school than in high school. If elementary school children "outgrew" the tests faster than high school students, then the lower gain-initial score correlations would be a by-product of not measuring high-scoring students' skills, but such a result would manifest itself as a ceiling effect on the elementary school tests. The Prospects tests used in the analysis in table 5 do not, however, show larger ceiling effects than the NELS tests.

following kindergarten. Yet learning during all other grades will not be separable into school year and summer components. This is a mistake because a number of studies, including the national Sustaining Effects study of the 1970s, suggest that the black-white gap widens over the course of schooling primarily because African American children gain less than white children during the summer (Entwisle and Alexander 1992, 1994; Hemenway et al. 1978; Heyns 1978, 1987; Klibanoff and Haggart 1981).[17] The only recent national data that speak to this question come from a nonrandom subset of Prospects students.[18] First, consider what we would conclude if we had measured black and white first graders' learning only at the beginning of first grade and the beginning of second grade. The first two columns of table 6 show that black and white children gain about the same skills in reading, vocabulary, and math over the course of a year. But dividing this yearly interval into learning that occurs during the winter (when school is in session) and learning that occurs during the summer (when school is not in session) yields a strikingly different picture of black-white differences in learning. The results show that, when school is in session, black first graders gain more reading and vocabulary skills than white first graders. Among students who start first grade with the same skills, blacks gain more vocabulary skills than, and about the same reading skills as, whites. But during the summer following first grade, blacks gain fewer reading and vocabulary skills than whites, both overall and among children who had similar skills at the end of first grade. The results for math show a similar but weaker pattern. These results resemble those from other studies (e.g., Heyns 1978, Entwisle and Alexander 1992, 1994).[19] Suppose we believe Phillips, Crouse, and Ralph's (1998) estimate that about half of the black-white gap is attributable to what occurs between first and twelfth grade. Without data on summer learning, we will never know how much of the black-

[17] The Sustaining Effects study was the first national longitudinal study to assess children's reading and math growth. In 1976, Sustaining Effects collected data on 120,000 students in over 300 public schools throughout the country (Carter 1983). It then followed a subsample of these students for three years, administering CTBS reading and math tests in the fall and spring of each school year as long as the students remained in the same schools (Won, Bear, and Hoepfner 1982).

[18] See appendix table A (page 132) for a comparison between this subsample and the larger cross-sectional sample. The subsample is quite advantaged relative to the original sample for reasons that remain a mystery to me.

[19] When Cooper and colleagues (1996) conducted a meta-analysis of the effect of summer vacation on test scores, they did not find consistent racial differences in summer learning. They did, however, find that middle-class children gained more in reading over the summer than lower-class children did. Cooper and colleagues explained that the lack of an effect of race was probably attributable to the fact that most of the studies in their meta-analytic sample partially controlled for family income before examining race differences.

Table 6. Black-White Differences in Reading, Vocabulary, and Math Growth on the CTBS in Prospects: School Year and Summer Comparisons

	Yearly Growth (fall, grade 1 to fall, grade 2)		Winter Growth (fall, grade 1 to spring, grade 1)		Summer Growth (spring, grade 1 to fall, grade 2)	
	Raw	Residualized	Raw	Residualized	Raw	Residualized
Reading Comprehension						
African American gain relative to European American	4.86	-4.97	19.59*	9.27	-14.73*	-20.47***
gain	(7.18)	(4.73)	(8.47)	(6.03)	(6.73)	(5.09)
Vocabulary						
African American gain relative to European American	13.67	5.46	24.38*	15.63*	-10.71**	-9.77*
gain	(10.22)	(5.21)	(10.68)	(6.37)	(3.66)	(3.32)
Math Concepts and Applications						
African American gain relative to European American	5.51	-3.13	11.26	8.62	-5.75	-18.75***
gain	(8.33)	(6.21)	(9.37)	(6.77)	(3.88)	(5.11)

NOTE: N=1,097. Numbers are unstandardized coefficients. Standard errors are in parentheses. Residual gain equations are errors-in-variables regressions that also control prior reading comprehension, vocabulary, and math concepts scores, corrected for measurement error using the reliabilities published in the Prospects Interim Report (1993). All equations are weighted using the 1993 weight, and all the standard errors are corrected for nonindependence within school districts. All equations also include gender and dummies for Asian American, Mexican American, Puerto Rican American, and other Hispanic students. See Phillips (1998) for more details on the sample and for results for other ethnic groups.

white gap is attributable to what occurs during the school year and how much is attributable to what occurs during the summer.

We typically design new surveys based on sparse and contradictory evidence, but the data on summer learning are too consistent to be ignored.[20]

[20] Nonetheless, the summer learning results still leave a number of important questions unanswered. We still do not know exactly how much of the widening of the black-white reading gap occurs during the summer as opposed to the school year. Nor do we know whether summers in early elementary school or later elementary school are most detrimental to black children's learning. Nor do we know how the relative contributions of families, schools, and neighborhoods to students' achievement change between the school year and summer vacation.

Summer vacations create a natural experiment that helps us begin to separate the effects of schooling from the effects of families and neighborhoods. Taking advantage of this experiment requires that we assess students twice a year: once in the fall and once in the spring. Testing students twice a year is expensive and imposes a considerable burden on students and schools, but testing students only once a year or once every two years is largely a waste of resources because it confounds what are potentially separable causes of differences in children's learning. A better alternative would be to randomly select multiple subsamples of students who would be tested twice a year. In the case of the ECLS-K, for example, one random subsample could be tested in the fall of first grade, another in the fall of second grade, another in the fall of third grade, and so on. This would minimize the burden to any particular student or school while maximizing our knowledge about how much of the learning gap arises during the summer.

Maximizing Measurement Variation within Data Sets

Another way to improve surveys of young children's academic growth is to increase the measurement variation within each survey. We can do this in the following two ways: by including multiple measures of the same skill and by including precise measures of each particular set of skills.

Multiple Measures of the Same Skill

Other than the Prospects survey, the only other contemporary national survey that has tested young children multiple times is the Children of the National Longitudinal Survey of Youth (CNLSY). Unlike school surveys, which have a grade-based sampling design, the CNLSY is age-based. The Bureau of Labor Statistics (BLS) first funded a survey of the mothers of these children in 1979, when the prospective mothers were 14- to 22-years-old. It has resurveyed them every year since then. In 1986, the BLS began collecting data on all the children who had been born to the original sample of mothers. These children, as well as all additional children born to the mothers, have been followed at two-year intervals since then.

The CNLSY administers math, reading, and vocabulary tests to all children who are at least five years old. These data show a small black-white *reading* gap (0.20 standard deviations) among 5-year-olds, but they show a large black-white *vocabulary* gap (0.98 standard deviations) among the very same children. If the CNLSY had administered more than one type of reading test and more than one type of vocabulary test to these children, we would be able to determine whether the stark difference in the size of these gaps reflects idiosyncrasies of the tests or a real phenomenon.

Precise Measures of Each Type of Skill

A survey can also enhance our understanding of ethnic differences in test scores by measuring a particular set of skills as precisely as possible. For example, the CTBS total math test administered by Prospects is composed of math computation and math concepts subtests. Analyzing these subtests separately reveals potentially important differences that we would miss if we only analyzed scores on the total math test. Table 7 shows ethnic differences in third graders' scores on the concepts and computation tests, as well as ethnic differences in gains on these tests between the third and sixth grades. The first panel shows that the gap between African American and white third graders is almost twice as large on the math concepts test (0.95 standard deviations) as on the math computation test (0.51 standard deviations). The same is true for the gap between whites and Mexican Americans. In contrast, Asian American third graders score at about the same level as white third graders on the math concepts test, but they score 0.65 standard deviations higher than whites on the math computation test.

The second panel in table 7 shows that African American and Mexican American children's *gains* also differ across the subtests. Although African American and Mexican American students gain about the same amount as whites on the math computation test, they gain more than whites on the math concepts test. These results illustrate why surveys need to administer tests that measure a wide range of skills and why analysts need to examine subtests separately, even when they sound relatively similar in name or content.

Maximizing Variation between Surveys

No matter how hard one works to perfect every survey, data sets inevitably end up with idiosyncrasies that can affect the analytic results. The best way to ensure that these biases do not affect our policy decisions is to collect as much data as possible, using different samples, survey designs, contractors, tests, technical review panels, and so on. Then, when we combine the results from these various studies, robust relationships should persist across the

Table 7. Comparison of Math Concepts Growth and Math Computation Growth: Evidence from Prospects Third Grade Cohort

	Math Concepts and Applications		**Math Computation**	
	1	**2**	**1**	**2**
Score in spring of third grade				
Asian American	4.65	8.32+	23.21***	24.06***
	(4.42)	(4.31)	(3.76)	(3.71)
	.10	*.17*	*.65*	*.68*
African American	-45.35***	-39.01***	-18.22***	-17.50***
	(2.66)	(2.60)	(2.28)	(2.25)
	-.95	*-.82*	*-.51*	*-.49*
Mexican American	-37.68***	-21.93***	-13.25***	-4.86+
	(3.07)	(3.10)	(2.67)	(2.71)
	-.79	*-.46*	*-.37*	*-.14*
Intercept	694.53***	679.25***	677.04***	672.32***
	(2.03)	(3.84)	(1.72)	(3.31)
Yearly growth between third and sixth grade				
Asian American	6.31***	4.65***	3.85**	3.73*
	(1.39)	(1.39)	(1.47)	(1.47)
	.36	*.27*	*.16*	*.15*
African American	2.75**	2.34**	.30	1.50
	(.87)	(.88)	(.91)	(.92)
	.16	*.14*	*.01*	*.06*
Score in spring of third grade				
Mexican American	4.66***	3.74**	-.48	.65
	(1.05)	(1.07)	(1.10)	(1.12)
	.27	.22	.02	.03
Grade	17.44***	19.34***	26.97***	25.06***
	(.80)	(1.40)	(.82)	(1.46)
Pseudo-R^2 for Intercept	.16	.25	.06	.15
Pseudo-R^2 for Slope	.04	.07	.03	.06

NOTE: N=4,550. Numbers are unstandardized coefficients. Standard errors are in parentheses. Italicized numbers are standardized coefficients for the intercept equation and proportions of the average gain for the slope equation. Equation 1 also includes gender and dummies for other ethnic groups on both the intercept and slope, and a nonlinear grade/age term on the slope. In addition, equation 2 includes dummies for mother's education, dummies for region, and dummies for urbanism on both the intercept and slope. All estimates come from weighted 2-level hierarchical models, with grade at level 1 and students at level 2. Standard errors and significance tests are corrected for clustering within districts by inflating the standard errors by the ratio of the standard errors produced by an unweighted 3-level model and an unweighted 2-level model. See Phillips (1998) for more details.

studies, despite their differences.[21] The best way to illustrate the importance of analyzing multiple surveys is to compare data from surveys that should, in theory, yield similar results.

For both the CNLSY and Prospects, I estimated growth models in which I predicted ethnic differences in students' initial test scores and learning rates.[22] The first panel of table 8 compares ethnic differences in initial reading scores in Prospects and the CNLSY. The CNLSY suggests that African American and white 5-year-olds have nearly identical reading recognition scores. (This finding resembles Entwisle and Alexander's BSS reading results for first graders in Baltimore.) Yet the results from Prospects show a 0.87 standard deviation gap in reading comprehension among first graders. Holding mothers' education, region of residence, and urbanism constant, this gap shrinks to 0.78 SDs.[23] (These results resemble those from the Sustaining Effects study.) If the CNLSY or Prospects had administered more than one reading test to their respondents, we would be able to determine whether the small black-white reading gap in the CNLSY is attributable to sample differences between the CNLSY and Prospects or to differences between the PIAT and CTBS tests.

[21] The same principle underlies meta-analysis. See Cook (1993) for a discussion of the principle of "heterogeneous irrelevancies." Of course, combining studies will not wash out biases that go in the same direction. For example, student mobility creates an upward bias in longitudinal studies because students who change schools or districts (and are therefore not retested) tend to have lower scores than students who stay in the same schools. This creates a major problem for researchers and policymakers who are often most interested in what happens to the most disadvantaged students.

[22] Children in the CNLSY first took Peabody Individual Achievement Tests (PIAT) in reading and math when they were 5 years old. I measured their learning rates in age-in-months. Children in Prospects first took Comprehensive Tests of Basic Skills (CTBS) in reading in the fall of first grade. I measured their growth rates in years. For details on the samples and analysis, see Phillips (1998).

[23] The contradiction between the reading results in Prospects and those in the CNLSY does not seem to be attributable to the fact that the CNLSY administered a reading recognition test, while Prospects administered a reading comprehension test. Equations using the CNLSY reading comprehension test as the dependent variable (not shown) produced results nearly identical to those for reading recognition. See Phillips (1998) for a discussion of whether the contradictory results are attributable to the psychometric properties of the different tests. She concludes that the kurtosis of the CNLSY tests probably does not account for the different results.

Table 8. Ethnic Differences in Reading Growth: Comparison of the CNLSY and Prospect

	CNLSY		Prospects	
	1	**2**	**1**	**2**
Reading score at age 5 (in fall, grade 1)				
African American	-.031 (.097)	.030 (.100)	-.868*** (.058)	-.776*** (.057)
Mexican American	-.552*** (.156)	-.430** (.164)	-.650*** (.087)	-.382*** (.088)
Linear reading growth in years				
African American	-.240*** (.024) *.121*	-.240*** (.024) *.120*	-.107** (.026) *.049*	-.071** (.025) *.034*
Mexican American	-.132*** (.041) *.070*	-.084+ (.048) *.042*	-.020 (.038) *.009*	.048 (.039) *.023*
Pseudo-R^2 for Intercept	.02	.12	.12	.20
Pseudo-R^2 for Slope	.10	.14	.04	.12

NOTE: To facilitate the comparison across data sets, all coefficients and standard errors are expressed as a proportion of the overall standard deviation of students' initial scores. In addition, italicized numbers in the gain equation express the gain as a proportion of the average gain in a particular data set. Moreover, the monthly gain coefficients and standard errors in the CNLSY have been multiplied by 12 to approximate the yearly gain interval in Prospects. Equation 1 includes gender and dummies for other ethnic groups on both the intercept and slope, and a nonlinear grade/age term on the slope. In addition, equation 2 includes dummies for mother's education, dummies for region, and dummies for urbanism on both the intercept and slope. All estimates come from weighted 2-level hierarchical models, with age/grade at level 1 and students at level 2. Standard errors and significance tests are corrected for the nonindependence of siblings (in the CNLSY) and for clustering within districts (in Prospects) by inflating the standard errors by the ratio of the standard errors produced by an unweighted 3-level model and an unweighted 2-level model. Prospects N=4,647; CNLSY N=2,153. See Phillips (1998) for more details.

Surprisingly, the CNLSY and Prospects results for Mexican Americans' initial reading skills are much more consistent.[24] The CNLSY suggests that Mexican American and white 5-year-olds' reading recognition skills differ by about 0.55 standard deviations. Prospects suggests that Mexican American and white first graders' reading comprehension skills differ by about 0.65 standard

[24] The similarity of these estimates is surprising in light of the fact that the CNLSY does not include children of recent immigrants. One would expect this sampling difference to affect comparisons of Mexican Americans' scores more than it affects comparisons of African Americans' scores, but that does not seem to be the case.

deviations. Among Prospects first graders whose mothers have the same amount of schooling and who live in the same region, the Mexican American-white reading comprehension gap shrinks by almost half—to 0.38 SDs. The reduction is not nearly so large in the CNLSY, but the gap after controlling mothers' education and region is remarkably similar (0.43 SDs).

The bottom panel of table 8 compares test score *growth* in the CNLSY and Prospects. Both the CNLSY and Prospects suggest that African American children learn fewer reading skills than white children during the early elementary school years, but the learning gap is less extreme in Prospects than in the CNLSY. In the CNLSY, African American children gain about 1 raw score point (over a fifth of the age-5 standard deviation) less than white children, each year. This is about 12 percent less than the average gain. Between the first and third grades, African American children in Prospects gain only 5.6 fewer points per year than whites, which is 5 percent less than the average gain and 11 percent of the first grade standard deviation.

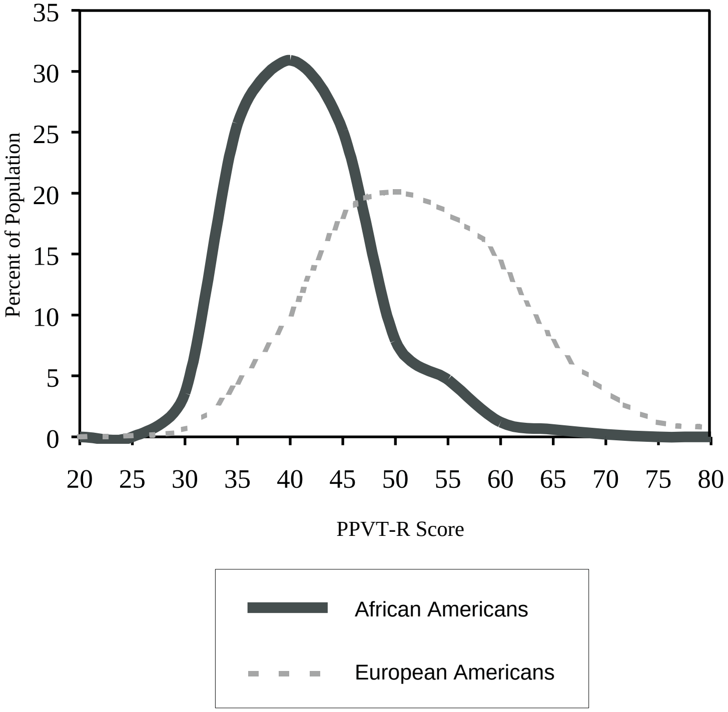
The CNLSY and Prospects also tell somewhat different stories about the relative reading trajectories of Mexican Americans and whites. In the CNLSY, Mexican American children gain about two-thirds of a raw score point (about 0.13 age-5 standard deviations) less than white children each year, which is equivalent to a gain of 7 percent less than average. This difference shrinks to a gain of 4 percent less than the average among white and Mexican American children whose mothers have the same amount of schooling and who live in the same region of the country. In contrast, Prospects suggests that young Mexican American children gain about the same reading skills as young white children, especially if we compare children whose mothers have the same amount of schooling and who live in the same region of the country.

The results in table 8 illustrate the difficulty of replicating results across surveys. They also serve as a cautionary tale to researchers who analyze data from a single source. Finally, these results underscore the problem of having only two contemporary national longitudinal data sets with which to describe elementary school children's academic development.

Studying Education Prior to Formal Schooling

Another reason we know so little about when ethnic differences arise and how they change with age is that NCES has, until recently, mostly ignored education that occurs outside of formal institutions. Although teachers have long argued that families exert the most influence on children's academic skills, few educational researchers study academic achievement before children enter elementary school. This habit hampers our understanding of ethnic differences in achievement because at least half of the black-white gap that exists at the end of twelfth grade can be traced to the gap that already existed at the beginning of first grade. Data from the CNLSY show that we can trace the vocabulary gap back to when black and white children are three years old (see figure 3). If

Figure 3. Vocabulary Scores for Black and White 3-Year-Olds, 1986–94



SOURCE: National Longitudinal Survey of Youth Child Data, 1986-94. Black N=507; white N=949. Figure is based on black and white 3-year-olds who took the Peabody Picture Vocabulary Test-Revised (PPVT-R). The scores shown are the standardized residuals, coded to a mean of 50 and a standard deviation of 10, from a weighted regression of children's raw scores on their age in months, age in months squared, and year-of-testing dummies. Lines are smoothed.

the CNLSY had measured infants' and toddlers' cognitive skills, we might be able to trace the gap back even farther.[25]

Focusing on family influences, as opposed to school influences, is not without its disadvantages, of course. Although the CNLSY does a better job than education surveys of measuring children's experiences outside of school, it does a considerably worse job of measuring their experiences inside school. Combining the advantages of a survey like Prospects with those of a survey like the CNLSY must become commonplace during the next century if we are to make any progress on the test score gap issue.

Fortunately, NCES is planning such a survey, in a joint effort with the National Center for Health Statistics (NCHS), the National Institutes for Child Health and Human Development (NICHD), the Administration for Children, Youth, and Families (ACYF), and the U.S. Department of Agriculture. The ECLS-Birth cohort study, targeted to begin in the year 2000, plans to follow a cohort of 6-month-olds through the end of first grade.

Hopefully, the designers of this study will learn from the strengths and limitations of the CNLSY. Important strengths of the CNLSY include sampling siblings in order to estimate family background effects, testing mothers' cognitive skills (testing mothers and fathers would be even better), and measuring parenting practices using both self-reports and interviewers' observations (observing parenting at more frequent intervals and using time diaries to collect parenting data would be even better). Limitations of the CNLSY include not trying to measure children's cognitive skills prior to age 3 and administering only one type of reading and vocabulary test to children.

Surveys versus Experiments

Studying elementary school students, assessing at least some of these students in the fall and spring of every school year, using multiple tests to measure students' reading and math skills, funding multiple surveys of chil-

[25] Because children's cognitive skills are moderately stable between infancy and early childhood, ethnic differences may be as well. Measures of infant habituation and recognition memory correlate 0.36, on average, with childhood IQ (see the meta-analysis by McCall and Carriger 1993). Thompson, Fagan, and Fulker (1991) also find that visual novelty preference at 5 to 7 months of age is associated with both IQ and achievement at age 3, and Dougherty and Haith (1997) find that visual anticipation and visual reaction time at 3.5 months are associated with IQ at age 4.

dren, and studying educational development before formal schooling begins—all would help improve our understanding of how ethnic differences in academic achievement change with age.[26] These design changes would also help us generate better theories about the causes of ethnic differences in academic skills.[27] Regardless of how we decide to design our national surveys, however, the best survey data will not tell us how to raise children's achievement—neither will

[26] An important methodological issue that I have not discussed here is the problem of choosing the correct metric with which to measure academic growth. Because the metric issue is so perplexing, almost all researchers simply use the particular test at their disposal, without questioning how the test's metric affects the results. For instance, when black and white children gain 50 points on a vocabulary test scored using IRT, we typically assume that black and white children *learned* the same amount of vocabulary over that interval. But even IRT-scored tests may not have interval-scale properties, which means that a gain from 250 to 300 may not be equivalent to a gain from 350 to 400. The only solution I see to the problem of determining whether gains from different points on a scale are equivalent is to associate a particular test with an outcome we want to predict (say, educational attainment or earnings), estimate the functional form of this relationship, and then use this functional form to assess the magnitude of gains. For example, if test scores are linearly related to years of schooling, then gains of 50 points can be considered equal, regardless of the starting point. If the *log* of scores is linearly related to years of schooling, however, then a gain of 50 points from a lower initial score is worth more than a gain of 50 points from a higher initial score. This "solution" is, of course, very unsatisfactory, because the functional form of the relationship between test scores and outcomes undoubtedly varies across outcomes.

[27] We should, however, question the assumption that large national studies should be preferred over multiple local studies. Ethnic differences in test scores vary across states, school districts, and schools. This means that national surveys probably mask much of the differential development in academic skills that occurs in particular school districts throughout the country. Scholars and policymakers should begin to debate the financial, political, and quality trade-offs between formally selecting nationally representative samples and purposefully selecting a large number of locally representative samples that are informally representative of the types of students to whom we would like to generalize nationally (see Cook [1993] on the logic of purposive sampling based on the "principle of proximal similarity"). Over the past few decades, we have learned the most about the correlates of young students' academic development from Entwisle and Alexander's BSS study. This may be because NCES never put the same effort into a national study of young children's achievement that it put into NELS. Or it may be that local surveys, run by university researchers who are better able to generate goodwill among local school administrators, teachers, and parents, are more effective than national surveys. In either case, a potential alternative to funding several large national surveys might be for NCES, in collaboration with private foundations and state and local governments, to support a large number of local longitudinal studies, on the condition that the procedures and measures be comparable enough to combine results across sites.

the best longitudinal methods.[28] Our ultimate goal in collecting data on student achievement is presumably to raise all children's achievement while reducing variation in achievement, not just to produce long descriptive reports or fill academic journals. If that is our goal, the standards that we need for assessing causality are much higher than survey data can satisfy.

The best way to learn what will reduce ethnic differences in achievement is to conduct randomized field trials.[29] Such studies could include programs designed by researchers, based on theories generated from nonexperimental data, as well as programs designed by teachers and administrators, based on programs that already seem to have worked on a small scale. The recent turn to natural experiments in economics and sociology may also help us begin to identify the causs of ethnic differences in academic achievement.[30]

We must also remember that it is not logically necessary to understand the causes of a social problem before successfully intervening to fix it. Instead of starting with the question "What causes the gap?" and hoping that the answers will lead to effective interventions, we may need to instead start with the question "How can we reduce the gap?" and then collect the kind of information (namely, experimental data) that will enable us to do just that.

[28] Methods for measuring longitudinal change and easy-to-use software packages for implementing these methods have proliferated over the past decade. Although these methods are often very helpful for describing longitudinal change and its correlates, they do not increase our ability to make correct causal inferences—nor are these methods always the best analytic choice. These new methods need to be subjected to cross-disciplinary discussions about their costs and benefits. The typical user neither understands the main advantages of multilevel models nor knows that other procedures (such as fixed effects models with a correction for the loss of degrees of freedom resulting from the intraclass correlation) have some of the same advantages without some of the disadvantages. See McCallum et al. (1997) for a review of multilevel methods for describing growth (including structural equation models) and Kreft (1996) for a discussion of the advantages and disadvantages of using random coefficient models.

[29] See Nave, Miech, and Mosteller (1998) for a review of the use of random field studies in education and for an argument in favor of funding more of them.

[30] When estimating the "effect" of educational processes on students' achievement gains, more frequent attention to correcting for measurement error in initial test scores would also help. In nonexperimental studies that find a positive effect of school or family characteristics on students' learning, the most frequent threat to validity is that these school or family characteristics mainly serve as proxies for initial skills that were imperfectly measured.

References

Alexander, K. L., and Entwisle, D. R. (1998). *Isolating the School's Contribution to Achievement: School Year Versus Summer Gains.* Paper presented at the annual meeting of the American Association for the Advancement of Science.

Ainsworth-Darnell, J. W., and Downey, D. B. (1998). Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance. *American Sociological Review 63*: 536–53.

Bloom, B. (1964). *Stability and Change in Human Characteristics*. New York: Wiley.

Carter, L. F. (1983). *A Study of Compensatory and Elementary Education: The Sustaining Effects Study. Final Report.* Santa Monica, CA: System Development Corporation.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Cook, P. J., and Ludwig, J. (1997). Weighing the Burden of "Acting White": Are There Racial Differences in Attitudes toward Education? *Journal of Policy Analysis and Management 16*(2): 656–78.

Cook, T. D. (1993). A Quasi-Sampling Theory of the Generalization of Causal Relationships. *New Directions for Program Evaluation 57*: 39–81.

Cooper, H., Nye, B., Charlton, K., Lindsay, J. and Greathouse, S. (1996, fall). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-analytic Review. *Review of Educational Research 66*: 227–268.

Cooper, H., and Hedges, L. (Eds.). (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Dougherty, T. M., and Haith, M. M. (1997, January). Infant Expectations and Reaction Time as Predictors of Childhood Speed of Processing and IQ. *Developmental Psychology 33*: 146–155.

Entwisle, D. R., and Alexander, K. L. (1988, May). Factors Affecting Achievement Test Scores and Marks of Black and White First Graders. *The Elementary School Journal 88*: 449–471.

Entwisle, D. R., and Alexander, K. L. (1990). Beginning School Math Competence: Minority and Majority Comparisons. *Child Development 61*: 454–71.

Entwisle, D. R., and Alexander, K. L. (1992, February). Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School. *American Sociological Review 57*: 72–84.

Entwisle, D. R., and Alexander, K. L. (1994, June). Winter Setback: The Racial Composition of Schools and Learning to Read. *American Sociological Review 59*: 446–460.

Fordham, S., and Ogbu, J. (1986). Black Students' School Success: Coping with the Burden of "Acting White." *Urban Review 18*(3): 176–206.

Hemenway, J. A., Wang, M., Kenoyer, C. E., Hoepfner, R., Bear, M. B., and Smith, G. (1978). *Report #9: The Measures and Variables in the Sustaining Effects Study.* Santa Monica, CA: System Development Corporation.

Heyns, B. (1978). *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Heyns, B. (1987). Schooling and Cognitive Development: Is There a Season for Learning? *Child Development 58*: 1151–1160.

Jencks, C., and Phillips, M. (1999). Aptitude or Achievement: Why Do Test Scores Predict Educational Attainment and Earnings? In S. E. Mayer and P. E. Peterson (Eds.), E*arning and Learning: How Schools Matter (*pp. 15–47). Washington, DC: Brookings Institution Press and Russell Sage Foundation.

Jencks, C., and Phillips, M. (1998). The Black-White Test Score Gap: An Introduction. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 1–51). Washington, DC: Brookings Institution Press and Russell Sage Foundation.

Kao, G., Tienda, M., and Schneider, B. (1996). Racial and Ethnic Variation in Academic Performance. *Research in Sociology of Education and Socialization 11*: 263–297.

Klibanoff, L. S., and Haggart, S. A. (1981). *Report #8: Summer Growth and the Effectiveness of Summer School*. Mountain View, CA: RMC Research Corporation.

Kreft, I. C. G. (1996). *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies.* Los Angeles, CA: California State University.

MacCallum, R. C., Kim, C., Malarkey, W. B., and Kiecold-Glaser, J. K. (1997). Studying Multivariate Change Using Multilevel Models and Latent Curve Models. *Multivariate Behavioral Research 32*(3): 215–253.

McCall, R. B., and Carriger, M. S. (1993, February). A Meta-analysis of Infant Habituation and Recognition Memory Performance as Predictors of Later IQ. *Child Development 64* (February): 57–79.

Nave, M., Meich, E. U., and Mosteller, F. (1998). *A Rare Design: The Role of Field Trials in Evaluating School Practices*. Paper presented at the American Academy of Arts and Sciences. Harvard University and the American Academy of Arts and Sciences. Cambridge, MA.

Ogbu, J. (1978). *Minority Education and Caste: The American System in Cross-cultural Perspective.* New York: Academic Press.

Phillips, M. (1998). *Early Inequalities: The Development of Ethnic Differences in Academic Achievement During Childhood*. Ph.D. dissertation, Northwestern University.

Phillips, M., Brooks-Gunn, J., Duncan, G., Klebanov, P., and Crane, J. (1998). Family Background, Parenting Practices, and Test Performance. In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 103–145). Washington, DC: Brookings Institution Press.

Phillips, M., Crouse, J., and Ralph, J. (1998). Does the Black-White Test Score Gap Widen After Children Enter School? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 229–272). Washington, DC: Brookings Institution Press.

Pindyck, R. S., and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts* (3rd ed.). New York: McGraw-Hill.

Rogosa, D. A., and Willett, J. B. (1985). Understanding Correlates of Change by Modeling Individual Differences in Growth. *Psychometrika 50* (2): 203–228.

Thompson, L. A., Fagan, J. F., and Fulker, D. W. (1991). Longitudinal Prediction of Specific Cognitive Abilities from Infant Novelty Preference. *Child Development 62*: 530–538.

Werts, C. E., and Hilton, T. L. (1977). Intellectual Status and Intellectual Growth, Again. *American Educational Research Journal 14*(2): 137–146.

Willett, J. B. (1988). Questions and Answers in the Measurement of Change. In E. Z. Rothkopf (Ed.), *Review of Research in Education* (pp. 345–422). Washington, DC: American Educational Research Association.

Won, E. Y. T., Bear, M. B., and Hoepfner, R. (1982, January). *Background, Schooling, and Achievement.* Technical Report 20 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills. Santa Monica, CA: System Development Corporation.

Table A. Descriptive Statistics for Prospects Summer Sample and Comparison with Cross-sectional Sample

	First Grade Cohort						
	Longitudinal			Cross-sectional			SD
Variables	Mean	SD	N	Mean	SD	N	Diff.
Reading comprehension	486.62	62.90	1,097	478.52	67.81	9,422	.12
Vocabulary	494.78	63.51	1,097	481.71	62.87	9,408	.21
Math concepts and applications	500.18	64.35	1,097	481.08	68.73	9,237	.28
Male	.51	.50	1,097	.51	.50	11,349	.00
Asian American	.01	.12	1,097	.03	.17	11,357	-.12
European American	.79	.40	1,097	.68	.46	11,357	.24
African American	.11	.31	1,097	.15	.36	11,357	-.11
Mexican American	.07	.25	1,097	.09	.29	11,357	-.07
Puerto Rican Am.	.01	.09	1,097	.01	.09	11,357	.00
Other Latino	.01	.10	1,097	.03	.18	11,357	-.11
Mom's educ in yrs.	13.29	2.00	1,097	12.92	2.10	10,529	.18
Mom is hs. dropout	.10	.30	1,097	.16	.37	10,529	-.16
Mom is hs. grad.	.23	.42	1,097	.28	.45	10,529	-.11
Mom has some coll.	.49	.50	1,097	.40	.49	10,529	.18
Mom is coll. grad.	.18	.38	1,097	.16	.36	10,529	.06
Live in north	.11	.31	1,097	.19	.39	11,357	-.21
Live in midwest	.27	.44	1,097	.18	.39	11,357	.23
Live in south	.48	.50	1,097	.38	.49	11,357	.20
Live in west	.15	.36	1,097	.24	.43	11,357	-.21
Live in urban area	.07	.26	1,097	.25	.43	11,357	-.42
Live in rural area	.49	.50	1,097	.38	.48	11,357	.23
Live in suburban area	.44	.5	1,097	.38	.48	11,357	.13

NOTE: Longitudinal sample includes children with valid original scale scores in fall of grade 1, spring of grade 1, and fall of grade 2, ethnicity, gender, mother's education, region, and urbanism, and is weighted with the spring 1993 weight. Cross-sectional sample includes all children selected in the base-year sample and is weighted by the 1991 weight. I used the original scale scores for these analyses because the fall of grade 2 adjusted scale scores are incorrect.

Certification Test Scores, Teacher Quality, and Student Achievement[1]

Ronald F. Ferguson with Jordana Brown
Malcolm Wiener Center for Social Policy
John F. Kennedy School of Government
Harvard University

Raising student achievement levels in primary and secondary schools is again a top national priority. Campaigning politicians at every level of government are promising to improve education, but many of the measures they are proposing have not been firmly established by research to be effective. This paper concerns standardized testing of teachers. In the constellation of measures that might contribute to achievement gains, certification testing of teachers is one of many, and not necessarily the most important.[2] Nonetheless, certification testing, especially for new teacher applicants, is now used in 44 states. In 1980, only three states tested teacher candidates. By 1990, the number had catapulted to 42. Given that so many states are now involved in this broad-based national experiment, a serious effort to learn from it seems warranted. This paper reviews evidence on the relationship of teachers' test scores to student achievement and frames some of the questions that a serious, nationally coordinated program of research might address over the next decade.

[1] Work on this paper has benefited from the financial support of the Annie E. Casey Foundation, the John D. and Catherine T. MacArthur Foundation, and the Rockefeller Foundation, through their support for the National Community Development Policy Analysis Network Project on Education and Youth in Community Development. One section of the text comes from another paper by this author, "Can Schools Narrow the Black-White Test Score Gap?" published recently in Christopher Jencks and Meredith Phillips, (Eds.), *The Black-White Test Score Gap* (1998). That section appears here with the permission of Brookings Institution Press.

[2] For example, achieving and maintaining a high level of quality in teacher education and professional development programs should be key elements in any strategy to improve teacher quality.

Measuring Teacher Effectiveness

The challenge of measuring teacher effectiveness applies to both new and incumbent teachers; and, in at least a few cases, standardized teacher testing has been used for both.[3] However, unlike new teachers, experienced teachers have other ways of demonstrating their effectiveness. It seems reasonable to argue that incumbents should be judged on what they do in the classroom, not on a test score. A commonsense procedure is to have expert observers rate teachers on their classroom practice. Alternatively, districts can judge teachers based on measures of student achievement, such as test-score gains. Hence, we have three indicators of current and potential teacher effectiveness: (1) teachers' test scores, (2) observers' ratings of teachers' professional classroom practice, and (3) students' achievement gains.

However, each of these measures has notable flaws. Regarding the first, certification tests may not measure those aspects of teacher skill that matter most. Regarding the second, observers may not rate teachers on the practices that matter most, or they may make mistakes in recording what they observe. Even the third, students' test-score gains, is an imperfect measure of what we really want to know: *the teacher's contribution* to producing the gains. Because other factors such as student, home, school, and community characteristics affect achievement as well, teachers deserve neither all of the credit for successes nor all of the blame for failures.

Ideal assessments of teacher quality would involve directly measuring what teachers contribute to student learning. Unfortunately, since such measures are infeasible, we must resort to various approximations—typically, one or more of the three types listed above or estimates that use them in multivariate statistical analyses. In a standard multivariate analysis, the dependent variable is the student test score. Explanatory variables include school, family, student, and community characteristics, and often a baseline value of the student test score. When a binary (i.e., 0,1) indicator variable for each individual teacher

[3] Two states, Arkansas and Texas, have tested incumbent teachers as well as those just entering the profession. Policymakers in Massachusetts are currently considering whether to test its current teachers, and other states may follow suit.

is included in such an analysis, its estimated regression coefficient is a measure of the teacher's contribution to the test score.[4]

Using this method of estimation, new findings for both Texas and Tennessee indicate that the teacher a student has in a particular year affects learning gains a great deal, not only for the current year, but for the next several years as well.[5] Estimated productivity differences among teachers in these new studies are large[6] and remind us how very important it is to select and retain the most effective teacher candidates.[7] When the data are adequate and the analysis is done appropriately, the technique that produced these new findings is probably the best that we can do at measuring the effectiveness of individual teachers.

[4] Under certain conditions, such a coefficient may be biased upward or downward. The rank order among teachers may even be distorted, such that some teachers appear more effective than others when they are not. The degree to which such analyses produce mistakes depends on such things as the completeness of the data used to control for confounding factors and the appropriateness of the specific techniques used for estimation.

[5] During the 1990s, teams of researchers have assembled large new longitudinal data sets for Texas and Tennessee. They include more students and more information than any previous compilation. These data permit researchers to follow tens of thousands of individual children's progress, including achievement gains associated with individual teachers, across several grade levels. For Texas, see Rivkin, Hanushek, and Kain (1998) and Kain (1998). For Tennessee, see Sanders and Rivers (1996) and Sanders, Saxton, and Horn (1998). There is an ongoing debate about how important it is to include additional student background variables when estimating the effects of teachers on students' test score gains. We agree with Kain (1998) that student background variables are important to include and that the teacher effects estimated by Sanders and his coauthors in the studies cited here might change (though probably not dramatically) if such controls were included.

[6] Earlier studies of this type have also identified large differences in teacher effectiveness. See Hanushek (1986, 1992).

[7] Of course, the challenge could be stated more broadly, to include affecting the size and composition of the teacher applicant pool. That might include considering a broader range of policy alternatives such as pay scales and career prep strategies for students in grades K–12 to attract them toward teaching careers. It might also include consideration of screening and hiring practices, particularly those that induce districts to hire the most effective applicants. For discussions of both salary-related issues and hiring practices, see Ballou and Podgursky (1997) and Murnane et. al. (1991). Another challenge is to help less effective teachers to improve their skills and knowledge. See, for example, Darling-Hammond (1997) for a discussion of these issues from the National Commission on Teaching and America's Future.

Unfortunately, this type of analysis is usually impractical. Few states or districts are anywhere close to possessing the data needed to implement it well on a large scale. Even for states that have the data and the capacity, the method does not work for judging new teacher candidates and others for whom there are not several years of appropriate data. Hence, standardized competency tests and other observational methods of judging professional practice offer more feasible alternatives for making judgments or predictions about professional effectiveness.

Whether competency testing is generally superior to observational teacher-rating methods is an open question that warrants more attention. However, it seems clear that competency testing is less expensive than observational methods when there are thousands of teachers to be assessed. It also seems clear that observational assessments of teaching *candidates*, conducted under contrived conditions or during student teaching, are likely to be highly variable in their quality. Given these cost advantages and the lack of alternative methods that are reliable and consistent, standardized teacher competency tests may be the best way of measuring teacher quality we have when thousands (or tens of thousands) of teachers need to be assessed.

Do Teachers' Scores Predict Student Achievement?

Most teacher certification exams are vulnerable to a variety of fair criticisms. For example, many have not been well validated and admittedly measure only a small fraction of the skills that make teachers effective. Yet evidence from studies that actually estimate relationships between students' and teachers' scores, including my own work reviewed below, suggests generally that teachers' test scores *do* help in predicting their students' achievement.

When I first encountered this topic in the late 1980s, it came as a great surprise to me that predictive validity in the relationship of teachers' to students' scores was not among the criteria for validating teacher certification exams.[8] States all over the nation were using such tests to screen candidates in and out of the profession, with no firm evidence that scores predicted teaching effectiveness! This was astonishing, especially since passing rates for nonwhite candidates were substantially lower than for whites. If teachers' scores

[8] This is still the case.

were poor predictors of student performance, then there was a disproportionate impact on nonwhite candidates that was probably illegal.[9] On the other hand, if it turned out that teachers' scores were important predictors, then there was a trade-off between the interests of low-scoring teacher candidates versus children's right to a quality education. As Bernard R. Gifford, an African American and Dean of the Graduate School of Education at the University of California wrote in 1985:

> If we do have a commitment to quality education for all, as part of our dedication to the principles of equality, then we will not change the requirements to fit the present median performance of minority applicants to teacher education programs. Rather, we will keep the desired performance level and provide the kinds of support and training that will make it possible for minority applicants to garner the learning and experience needed to pass the examinations . . .
>
> To put forth the argument that minority youngsters, the most disadvantaged of the poor, and the least able to emancipate themselves from their impoverished surroundings, should be taught by our less-than-best teachers is to pervert the nature of justice. As admirable and important as is the goal of increasing the ranks of minority teachers, this objective must not be put before the more fundamental objective of securing good teaching for those who need it the most.[10]

In this passage, Gifford assumes that the exams measure skills that matter for predicting teacher effectiveness. Many others have disagreed with this assumption.[11] The lack of a well-organized body of evidence on how teacher

[9] Pressman and Gartner (1986, 11) went so far as to assert, "There is no evidence that what is being tested relates to the selection of persons who will be effective teachers." Their article also discusses some of the legal challenges that tried to stop the use of certification exams.

[10] See Gifford (1985, 61).

[11] For example, see Pressman and Gartner (1986) for a very skeptical discussion about competency testing.

characteristics (including race and test scores) relate to teacher effectiveness has fostered confusion in both scholarly and public policy discourse.[12]

Robert Greenwald, Larry Hedges, and Richard Laine (1996) pool findings from all published education production function studies that fit their criteria for inclusion.[13] Because research has not focused on teachers' scores, only 10 of the studies include measures of teacher test scores among the predictors of student achievement. Among these 10, most are over a decade old and use data from the 1960s and 1970s. The 10 studies include 24 independent coefficients measuring the relationship of teachers' scores to their students' standardized achievement scores. Among the 24 coefficients, 21 are positive, and only 3 are negative. Among the 21 positive coefficients, 12 are statistically significant at the .05 level. In their statistical meta-analysis, the authors address whether this pattern of coefficients across all of the studies might result purely by chance if there is no relationship in general between students' and teachers' scores. Their answer is unambiguously no. Some aspects of the Greenwald, Hedges, and Laine analysis have been challenged on methodological grounds, but the findings regarding teacher test scores have not; other methods of summarizing the literature would lead to the same general conclusion.[14]

Simply stated, even though the number of studies is relatively small, it appears generally that teachers who score higher on tests produce students who do also. Let me emphasize that no one characteristic of a teacher is a

[12] For example, see the response by Cizek (1995) to King (1993) in the *Review of Educational Research.*

[13] These standards were (1) the study was in a scholarly publication (e.g., a refereed journal or a book); (2) the data were from schools in the United States; (3) the outcome measure was some form of academic achievement; (4) the level of aggregation was at the level of the school district or a smaller unit; (5) the model controlled for socioeconomic characteristics or for prior performance levels; and (6) each equation was stochastically independent of others, such that only one of several equations from a study that used the same students but different outcome measures (e.g., math scores in one equation but reading scores in another) was kept for the Greenwald, Hedges, and Laine analysis.

[14] A response by Hanushek (1996) disputes the Greenwald, Hedges, and Laine interpretation of the evidence regarding school resources, especially class size. Hanushek does not dispute the findings regarding teachers' test scores, however. This is apparently because the vote-counting method that Hanushek tends to prefer would produce the same basic conclusion about teachers' scores. There is no clear winner of the debate regarding the methodological issues, because each side is correct if its favored assumptions are true, and there is no neutral way to test the validity of the assumptions.

totally reliable predictor of his or her performance. Nor are most teachers uniformly strong or weak in every subject or with all types of students. Nevertheless, until we develop more and better research to test it more completely and rigorously, my judgment is that a positive causal relationship between students' and teachers' scores should be the working assumption among policymakers.[15] Below, I present some more detailed evidence that supports this judgement.

Evidence from Texas and Alabama[16]

During the late 1980s, I constructed a data set for about 900 districts in Texas. Data were aggregated to the district level, and variables included teachers' scores on the Texas Examination of Current Administrators and Teachers (TECAT). Texas required all of its teachers to pass the TECAT or relinquish their jobs in 1986.[17] The test was essentially a reading, vocabulary, and language skills test, geared to about an eleventh grade level of difficulty. Some teachers had to retake it, but most eventually passed. Controlling statistically for a host of school and community characteristics, I found that district-average TECAT scores were strong predictors of why some school districts had higher student reading and math scores and larger year-to-year gains (Ferguson 1991). Thus, at least for Texas, a certification test that measured no specific teaching skills and that challenged teachers at only an eleventh grade level of difficulty seemed to distinguish among levels of teacher effectiveness. Later, Helen Ladd and I found similar patterns for Alabama, using teachers' college entrance exam (i.e., ACT) scores from when they applied to college.[18]

[15] One hypothesis is that teachers who score high on tests are good at teaching students to do well on tests or that they place greater emphasis on test taking skills, and that is why their students score higher. By this hypothesis, test score differences overstate "true" differences in how much children have learned. I have found no research that tries to test the validity of this hypothesis or to gauge the magnitude of any associated overstatement of differences in learning.

[16] This section draws extensively from an earlier paper of mine (Ferguson 1998) with the permission of Brookings Institution Press.

[17] Not many lost their jobs in Texas because with second and third chances most passed (I do not know the details for Arkansas). For the story of what happened in Texas, see Shepard and Kreitzer (1987).

[18] See Ferguson and Ladd (1996).

Below I present some new estimates, using the Texas and Alabama data and adding a few distinctions to my previous work.[19] I review evidence from Texas in the 1980s showing that teachers' scores were lower where larger percentages of students were black or Hispanic.[20] I also present evidence that test score gaps among teachers contributed to the black-white test score gap among students. Further, I use data from Alabama to show that when certification testing reduces entry into teaching by people with weak basic skills, it narrows the skill gap between new black and white teachers. Finally, I suggest that because rejected candidates would probably have taught disproportionately in black districts, initial certification testing for teachers is probably helping to narrow the test score gap between black and white students in Alabama.

Teachers' Scores and Students' Race-Ethnicity

Texas tested all of its teachers in 1986 using the TECAT. Black teachers had lower scores than white teachers by more than a standard deviation, and black teachers were more likely than white teachers to teach in districts with many black students.[21] (See column 1 of table 1.) Moreover, white teachers who taught in more heavily black and Hispanic districts tended to have lower scores than other white teachers. (See column 2 of table 1.) In Texas, and certainly in other places too, attracting and retaining talented people with strong skills to teach in the districts where black and Hispanic students are heavily represented is part of the unfinished business of equalizing educational opportunity.

[19] Specifically, the earlier paper (Ferguson 1991) did not have separate scores for elementary and high school teachers. Now, having both elementary and high school teachers' scores provides the basis for testing whether the difference between third-to-fifth and ninth-to-eleventh grade gains is a function of the difference between elementary and high school teachers' scores. See below.

[20] For more statistical estimates using these data, see Ferguson (1991). Kain is also currently assembling a large data set for Texas with which to study student performance at the individual level. See Kain (1995) and Kain and Singleton (1996) for two early papers from the project.

[21] This standard deviation is for the statewide distribution of scores among individual teachers.

Table 1. Effects of Percent Minority on Teachers' Test Scores and of Teacher Test Scores on Student Mathematics Achievement, Texas (District-level data; *t*-statistics in parentheses)

Dependent Variables:							
	Teachers' TECAT Scores		Math Scores for Students in 1988				
Explanatory variable	**Black Teachers**	**White Teachers**	**Fifth Grade**		**Eleventh Grade**		**Difference in Gains: 9th–11th minus 3rd–5th**
Column	1	2	3	4	5	6	7
TECAT Scores							
HS Teachers	...	...	...	...	...	0.128	0.164
						(3.83)	(2.09)
Elem. Teachers	...	...	...	0.146	...	...	-0.179
				(2.96)			(2.13)
Percent Minority Students:							
Black	-0.031	-0.014	-0.006	-0.0004	-0.019	-0.015	-0.012
	(3.52)	(9.58)	(1.81)	(0.12)	(7.83)	(5.70)	(2.07)
Hispanic	-0.013	-0.010	-0.008	-0.007	-0.014	-0.013	-0.007
	(2.93)	(14.03)	(3.35)	(2.78)	(8.33)	(7.73)	(1.91)
Mathematics scores, 1986							
Third Grade	...	...	0.394	0.367	...	...	...
			(12.08)	(11.41)			
Fifth Grade	...	...	...	...	0.455	0.438	...
					(16.92)	(16.15)	

Table 1. Effects of Percent Minority on Teachers' Test Scores and of Teacher Test Scores on Student Mathematics Achievement, Texas (District-level data; *t*-statistics in parentheses) (continued)

Additional Var's*	a	a	b	b	b	b	b
N	386	919	884	884	853	853	849
Adj. R-squared	0.03	0.19	0.46	0.47	0.72	0.73	0.04

NOTE: Observations are for districts, and data are district-level averages. Teachers' TECAT scores and students' math scores are measured in standard deviations of the mean scores among districts. For both TECAT scores and students' scores, this standard deviation from the distribution of district-level means is roughly one-third the standard deviation of the statewide distribution of scores among individuals. Each observation in each regression is weighted by the square root of student enrollment.

*Additional Variables: (a) Constant term is only control variable. (b) Control variables: teachers per student; percent of teachers with master's degrees; percent of teachers with 5 or more years experience; percent of adult population with more than high school education; percent of adult population with high school education; log per capita income; percent of children in poverty; percent of female-headed families; percent in public schools; percent of migrant farmworker students; percent English-as-a-second language; and indicator variables for city, suburb, town, nonmetro city, rural, Mexican border high-poverty district.

Teachers' Scores Help Predict Racial Test Score Gaps

Estimates using the Texas data and standard econometric specifications for education production functions show that TECAT scores are important predictors of students' math scores. (See columns 4 and 6 of table 1.)[22] In addition, teachers' scores help to explain why average math scores are lower in districts where larger percentages of students are black.[23] However, we cannot be sure that teachers' test scores affect students' test scores, because teachers' scores might merely be standing in for some omitted variables that are correlated with both teachers' and students' scores. Fortunately, separate scores for elementary and high school teachers allow me to circumvent this problem.[24] I compare high school gains to elementary school gains in the same district and ask whether the difference in high school and elementary school gains is larger in districts where the TECAT gap between high school and elementary school teachers is larger.[25] Using this approach, a change of one standard deviation in teachers' TECAT scores predicts a change of 0.17 standard deviation in students' scores over the course of two years.[26]

[22] Table 1 shows regression results where the dependent variable is the math score in 1988 for fifth grade (columns 3 and 4) or eleventh grade (columns 5 and 6). Two of the four columns include teachers' scores among the explanatory variables. All four columns include math scores for the same cohort from 1986. Including earlier scores for the same cohort among the explanatory variables is a standard way of estimating gains in achievement since the earlier date.

All of the regressions reported in table 1 are weighted by the square roots of district enrollment. This is a standard fix-up for heteroskedasticity in cases where data are means from samples of different sizes. Houston and Dallas are not included in the analysis because of a poorly conceived decision that I made when constructing the data set several years ago. For a detailed description of the data, see Ferguson (1991).

[23] Compare the coefficient on "percent black among students" from column 3 with that in column 4; and compare the coefficient in column 5 with that in column 6. Note that percents black and Hispanic are on a scale of 0 to 100.

[24] This of course assumes that unmeasured factors affecting *differences* between elementary and secondary students' test score gains are not correlated positively with *differences* between elementary and secondary teachers' scores.

[25] The dependent variable in column 7 of table 8 is the difference between two differences: (a) the district's mean high-school gain between the ninth and the eleventh grades, minus (b) the district's mean math score gain between the third and the fifth grades. Elementary and high school teachers' TECAT scores are included as separate variables.

[26] Here, 0.17 is the average of 0.164 (the coefficient on high school teachers' scores) and 0.179 (the absolute value of the coefficient on elementary school teachers' scores) from column 7 of table 8.

If the impact of skilled teachers is important *and accumulates*, then unusually high (or low) average TECAT scores for an entire district should help to pull up (or down) students' scores, and this impact should become more starkly apparent, the longer children are in school. For example, among districts where students do poorly in the early years of elementary school, districts where TECAT scores are unusually high should achieve much higher student scores by the end of high school than districts where TECAT scores are unusually low. To test this, I selected four sets of districts for comparison: districts with unusually high TECAT scores but low first- and third grade math scores (N=3); districts with unusually high TECAT scores and high first- and third grade math scores (N=37); districts with unusually low TECAT scores and low first- and third grade math scores (N=25); and districts with unusually low TECAT scores and high first- and third grade math scores (N=4).[27]

For each of the four sets of districts, figure 1 graphs the district-average math score for grades 1, 3, 5, 7, 9 and 11 for the 1985–86 school year.[28] Compare the patterns for districts that have similar teachers' scores. The dashed lines are districts where teachers' scores are more than a standard deviation above the statewide mean. Even though they start at opposite extremes for first- and third grade scores, the two have converged completely by the 11th grade. The solid lines are districts where teachers' scores are more than a standard deviation below the statewide mean. Here too, students' scores have converged by the eleventh grade, but at a far lower level.

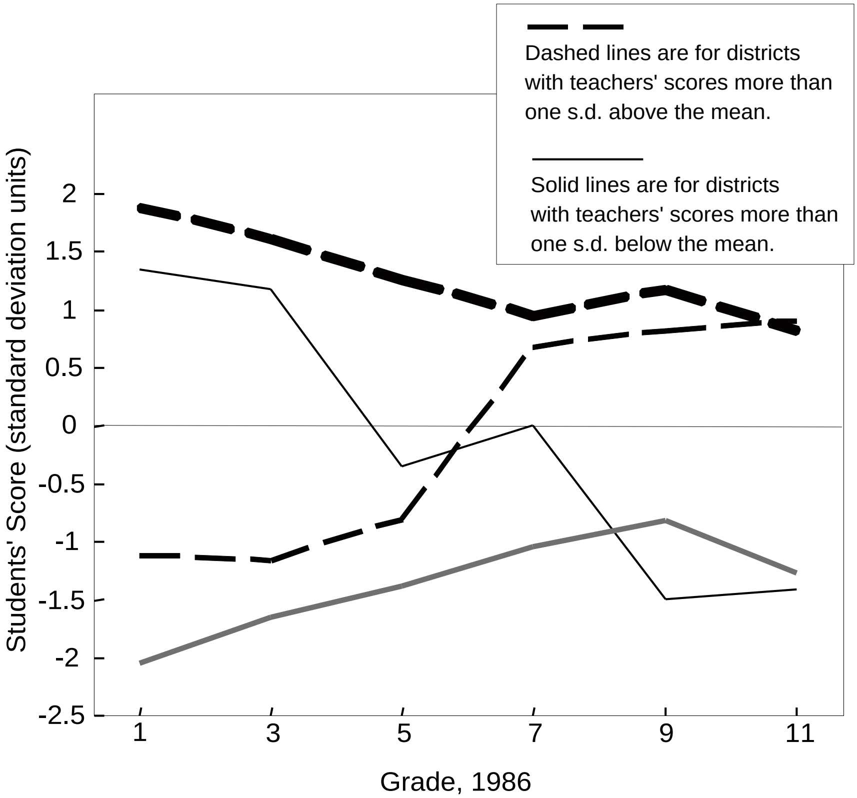
Figure 1 is not absolute proof of causation, but it is exactly what one would expect under the assumption that teachers' measured skills are important determinants of students' scores. Also, the magnitude of the change in

[27] I define "unusually" high (or low) to be a district-average TECAT score of more than one standard deviation above (or below) the statewide mean, where the relevant standard deviation is that among district-level means. Districts with low first and third grade math scores are those where math scores are more than a half standard deviation below the statewide mean for both years. Here too, the relevant standard deviation is that among district-level means. For both students' and teachers' scores, the ratio of statewide individual-level to district-level standard deviations in these data is 3 to 1.

Districts with high-scoring teachers and low-scoring students or low-scoring teachers and high-scoring students are rare. This is why, from roughly 900 districts, I could identify only a few, as indicated in the text and in the note to figure 1.

[28] A diagram for students' reading scores (not shown) follows the same general pattern, as do similar graphs using data for Alabama, albeit less dramatically.

Figure 1. Effect of Teachers' Test Scores on District-Average Mathematics Test Scores across Grades, Texas, Selected Districts, 1985–86



SOURCE: Author’s calculations based on data obtained from the Texas Education Agency.

NOTE: Sample comprises three districts with unusually high teacher scores on the Texas Examination of Current Administrators and Teachers and unusually low scores on first and third grade mathematics achievement tests; four districts with low teacher scores and high first and third grade student scores; 37 districts with high scores for both teachers and students; and 25 districts with low scores for both teachers and students. For TECAT scores, “high” and “low” mean one standard deviation or more above and below, respectively, the Texas mean; for mathematics scores, the respective criteria are 0.50 standard deviations above and below the Texas mean. Standard deviations for both teachers’ and students’ scores are from the distribution of district-level means. In each case, the ratio of this standard deviation to that for individuals statewide is 3 to 1.

figure 1 from elementary through high school is almost exactly what one would predict using the regression estimates from column 7 of table 1. Specifically, for two districts starting with equal student scores, but teachers' scores separated by two standard deviations, over 10 years the difference in student scores would accumulate to 1.70 standard deviations.[29] This is a large effect.[30]

Certification Testing Probably Narrows the Black-White Test Score Gap

Relying less on evidence from research than on their own judgment, policymakers in 43 states had enacted some form of initial competency testing for teachers as of 1996.[31] Thirty-nine states include a test of basic reading and (sometimes) math skills. This is usually supplemented by an additional test of professional knowledge, such as the National Teachers Exam (NTE, now called PRAXIS), which is (as of 1996) used in 21 states.

Initial certification testing restricts entry into the teaching profession. Figure 2 shows the effect of certification testing on the mix of people who became teachers after Alabama began requiring certification tests in 1981. The data are from teachers' ACT scores at the time they applied to college.[32] After certification testing began, the test score gap between new black and white teachers fell sharply. Since districts in Alabama that have more black students also have more black teachers,[33] a change that increases the average level of skill among incoming black teachers should disproportionately benefit black children. If this pattern recurs in other states, as seems likely, we should find that black children's scores improve more than white children's scores after states

[29] 1.70=0.17 x 2 s.d. x 5 two-year intervals.

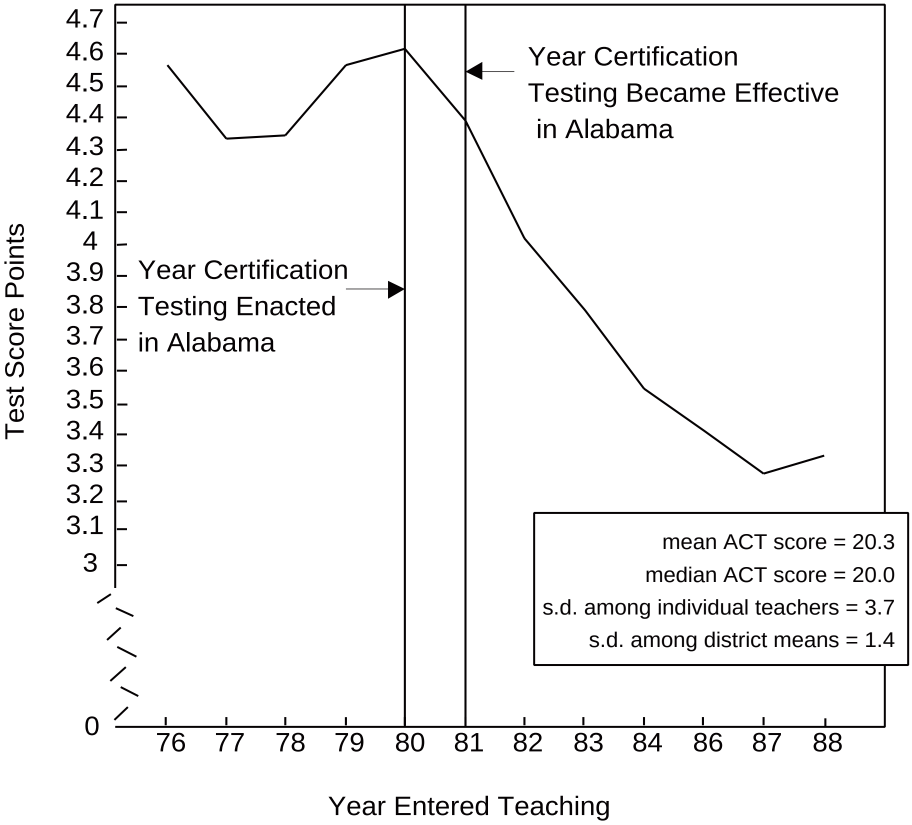
[30] This is not simply regression to the mean for student scores. Note that there are *two* sets of districts whose student scores are far below the mean as of the first and third grades. Only the districts with high teacher scores have student scores above the mean by the end of high school. Scores do regress toward the mean for the districts with low teacher scores, but these student scores nevertheless remain substantially below the mean. A similar set of statements applies to the districts whose first and third grade scores are above the mean.

[31] See U.S. Department of Education (1996). Table 154.

[32] See Ferguson and Ladd (1996) for more detail on the ACT data for Alabama.

[33] The simple correlation of "percent black among students" and "percent black among teachers" is 0.91 among 129 districts in Alabama.

Figure 2. Difference between Mean College Entrance Exam Scores of White and Black Teachers by Year of Entry into the Profession, Alabama, 1976–88



SOURCE: Author’s calculations, unpublished data. ACT scores are from teachers’ college entrance exams and are not associated with any certification exams that they may have taken.

implement certification testing for teachers (but we should expect some improvement even for whites).

Twenty-five years ago, working with data from the 1966 Coleman report, David Armor wrote:

> Even though black teachers’ formal training seems as extensive as that of white teachers, if not more so, their verbal scores indicate that they have far less academic achievement. It is especially

> ironic, when schools are concerned with raising black student achievement, that the black teachers who have the major responsibility for it suffer from the same disadvantage as their students.[34]

Once certification testing began in earnest after 1980, passing rates for black applicants in states across the nation were sometimes half those for whites.[35] Certainly, some black teachers who failed would have become good teachers. However, the relevant policy question is whether students on average are better off with the policy in place. I think the answer is yes.[36] However, truly definitive answers would require better data, developed and utilized in a multistate, longitudinal program of research.

We Need Better Data

Testing whether teachers' test scores or observers' ratings are good predictors of professional effectiveness is not a simple process. Even when there is agreement that gains in pupils' test scores should be the primary measure of professional output, a number of statistical assumptions must hold in order for studies to produce reliable estimates of how well teachers' scores (or ratings) measure their effectiveness. Problems associated with measurement error, incorrect functional forms, omitted variable bias, simultaneity bias, and reverse causation plague this type of analysis. Authors in the education production function literature over the last few decades have encountered these problems routinely (but seldom overcome them). In addition, during the 1990s, statisticians have emphasized the importance of hierarchical models to distinguish student-level from school-level from district-level effects of explanatory variables.[37]

[34] See Armor (1972).

[35] Quoting numbers from Anrig (1986), Irvine (1990, p. 39) presents the following numbers: "In California, the passing rate for white test-takers was 76 percent, but 26 percent for blacks; in Georgia, 87 percent of whites passed the test on the first try, while only 34 percent of blacks did; in Oklahoma, there was a 79 percent pass rate for whites and 48 percent for blacks; in Florida, an 83 percent pass rate for whites, 35 percent for blacks; in Louisiana, 78 percent for whites, 15 percent for blacks; on the NTE Core Battery, 94 percent of whites passed, compared with 48 percent of blacks."

[36] Available estimates suggest that the impact of teachers' scores on students' scores does not depend on the race of the teacher. Ehrenberg and Brewer find this using the verbal skills test from Coleman (1966). I also find it in unpublished results using data for Texas.

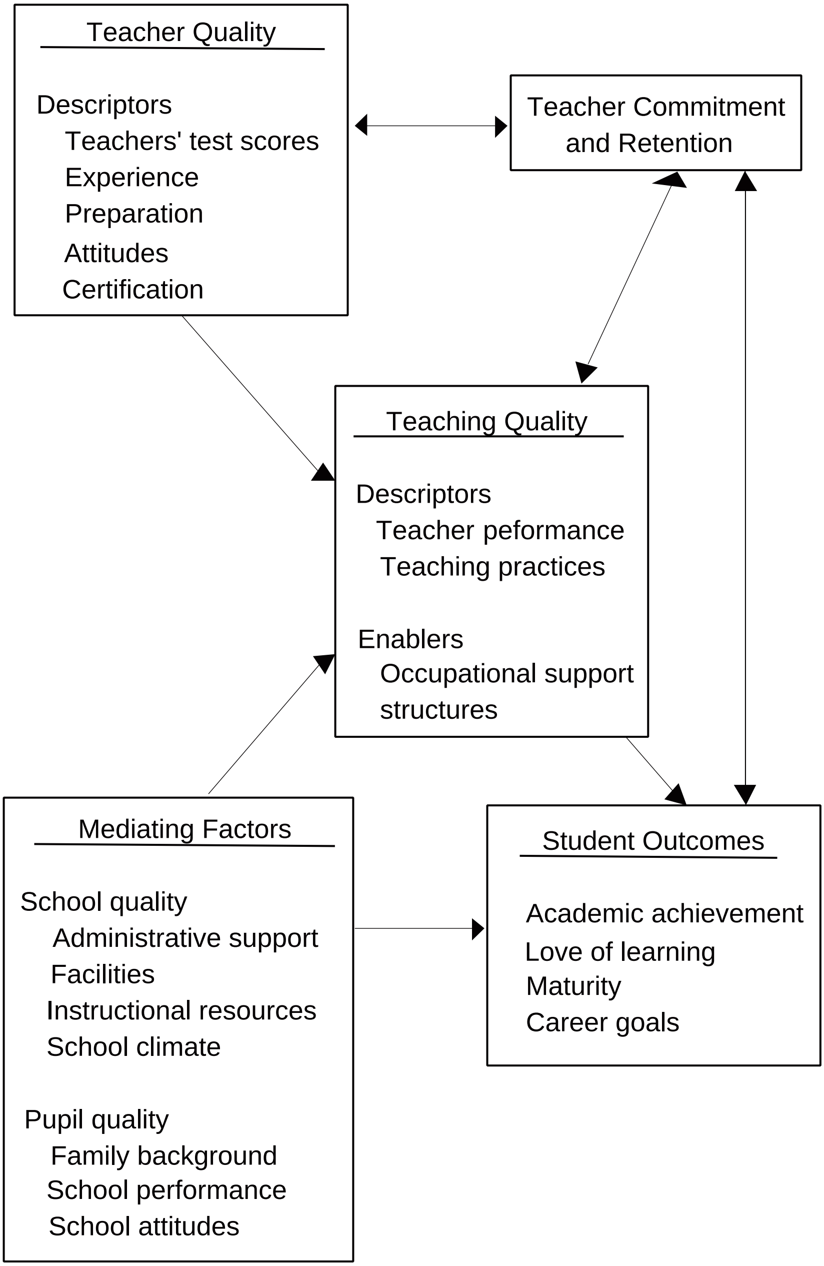
[37] Other papers in this volume elaborate the advantages of multilevel modeling.

Teacher quality measures present all of the standard statistical problems listed in the paragraph above: teacher quality, including teachers' test scores, are measured with error; student quality in a school or district can affect which teachers choose to apply there, creating reverse causation from student performance to teacher quality; particular measures of teacher quality may matter more or less depending on other variables, such as class size, so simple linear models that ignore interactions may produce misleading results; correlations between teacher quality and other inputs such as parental effectiveness can produce biased estimates for the effect of teacher quality if parental variables are omitted from the analysis or measured with considerable error. Further, most studies lack the type of data necessary for sorting out the issues that the advocates of multilevel estimation emphasize.

Even when measuring the effect of teacher quality on student outcomes is the only goal, data requirements can be vast. It is difficult to emphasize enough that teaching is a complex process in which context matters. Helping students to achieve academic success, love of learning, maturity, or career success involves far more that high certification test scores. Indeed, Darling-Hammond and Hudson (1989) distinguish *teacher* quality (e.g., certification test scores, experience, preparation, attitudes, aptitudes) from *teaching* quality (i.e., performance in the classroom). Further, they point out that how effectively both teacher and teaching quality translate into student outcomes depends on characteristics of schools, students, and families. (See figure 3 for a summary picture.)

Since no analyst will ever achieve a fully specified statistical model of this process or have the ideal data for solving *all* the statistical problems listed above, we will never reach perfection. We can, however, do better than we have. Researchers seem to agree on at least two points. First, we need more random assignment experiments to test hypotheses about the productivity of schooling inputs such as small-versus-large classes or high-versus-low teacher test scores. Second, because random assignment studies are sometimes impractical, we need more student-level longitudinal data sets that include good measures of child, teacher, family, classroom, school, and community characteristics. Further, no matter how carefully we assemble longitudinal data, the possibility that results are driven, for example, by omitted variable bias, will always make the findings from individual studies less than definitive. Similarly, findings from a single random assignment study may depend on idiosyncratic conditions that are not maintained at other times and places. Hence,

Figure 3. A Model of Teacher-Quality Effects (Adapted from Darling-Hammond and Hudson, 1989)



for both random assignment experiments and statistical studies using longitudinal data, we need replication across multiple independent analyses. There has been no organized program of education research to test whether standard measures of teacher quality are reliable predictors of student learning.

Future Research Involving Teachers' Scores

The following are five sets of issues and questions that a future program of research could usefully address about measures of teacher quality, all involving teacher test scores.

1. **Best Practices**. Do teachers' own test scores predict whether they use practices in the classroom that researchers have classified as most effective?[38] Or is the apparent relationship between teachers' scores and student performance measuring something subtler than so-called best practices?[39]
2. **Fixed Effects**. Using fixed-effects specifications, researchers can estimate which teachers consistently over the years produce greater learning gains, as measured by changes in their students' standardized test scores.[40] Do teachers' own scores predict the teacher-effects that these studies estimate? If we put measures of effectiveness based on observer ratings into the same equations that include teachers' scores, does the predictive power of teachers' scores remain unshaken?
3. **Generalizability and Fairness**. Are teachers' scores equally accurate predictors for teachers with different characteristics (e.g., different ethnicities, different training, and so on)?
4. **Effects of Other Inputs**. Teachers' scores can also be control variables. More and better teacher test score data would provide better statistical controls for estimating the effects of other variables such as experience, masters degrees, class size, or even parents' education.

[38] For literature reviews regarding teaching quality and best practices, see Brophy (1986), Doyle (1986), Darling-Hammond and Hudson (1989), and Porter and Brophy (1988).

[39] For example, it could be that teachers with more skills are those with the better judgment — for example, those who *depart* from generally effective practices at precisely those times that the practices would not be effective.

[40] See Rivkin, Hanushek, and Kain (1998) and Sanders and Rivers (1996).

5. **Allocation of Teacher Quality**. Attracting more strong teaching candidates and having them teach where they are needed most is important. Who gets the best teachers and why? It would be useful to know the degree to which salaries and other factors are important predictors of where high-scoring teachers end up teaching (e.g., which grades, schools, tracks, districts).

Certainly, the list could be longer. However, a serious program of research that made important progress on these five sets of issues would be a big step forward.

The National Center for Education Statistics (NCES) could help in the following ways:

1. Information about teachers, for example, the college that the teacher attended for BA and MA degrees, could be added to the teacher surveys that accompany NCES student surveys.
2. NCES can convene and coordinate state-level researchers who are constructing longitudinal student-level data sets that include (or can include) teachers' scores and other teacher characteristics.
3. NCES can encourage the Educational Testing Service and other test makers to work with states to validate teacher exams as predictors of student performance.
4. NCES should increase the number of students sampled per teacher in longitudinal NCES data series. NCES could also facilitate matching of its data with state-level data for teachers and students.

None of these will be easy, but each would be helpful.

Conclusion

Difficulty talking in public about racial and ethnic differences in test score patterns is probably a major factor in why the nation has not addressed these issues with the seriousness that they deserve. This challenge needs to be confronted. As I write, public officials in the state of Massachusetts are debating whether to test incumbent teachers for recertification. The basis of their interest in testing is the belief that certain elements of core knowledge are foundations for professional practice. Any teacher who lacks this knowledge cannot, the theory goes, be an effective teacher. The bulk of the evidence that we have suggests that teachers' scores on even the most rudimentary of basic skills exams—for example, the 30-item test in the Coleman study or the TECAT test in Texas—can be statistically significant predictors of how much students will learn. Regarding whether to screen teacher candidates using such exams, I am inclined to give the benefit of the doubt to students, which for me means endorsing the continued use (*and ongoing improvement*) of certification exams. As Bernard Gifford suggested, it is better to work on raising the skills of teaching candidates who might otherwise fail than to lower the standards that teaching candidates are expected to meet, and thereby to raise the risk that children will receive poor schooling.

On the other hand, our knowledge is far from definitive and very incomplete. Current certification exams produce an unknown number of mistakes that cause individuals to suffer unfairly. Some candidates who rate high on dimensions that tests do not measure and who would have been good teachers fail certification exams and never become teachers. Conversely, some are "false positives" who pass the exams but may fail in the classroom. Nonwhite candidates are probably over-represented among the false negatives who fail the exams but would have been good teachers.[41] At the same time, nonwhite children are probably over-represented among beneficiaries. This is because more of the people who fail, and would *not* have been good teachers, would probably have shown up to teach in classrooms where nonwhite children are over-represented. We may never know for sure. Nonetheless, I believe that if we had better data, a greater willingness to debate hard questions, and a targeted program of research, we would find ways to be more nearly fair in selecting among teaching candidates and ultimately more effective in helping those hired to become good teachers.

[41] See the discussion in Jencks and Phillips (1998, 77). Assume that a test score is the only basis for selecting people into a job, such as teaching. Also assume that black candidates, on average, have lower average scores than whites but are more similar to whites on other skills that affect teaching quality. Jencks explains why a larger percentage of blacks will be excluded than whites, among those people who would have performed well if hired (or do perform well).

References

Anrig, G. R. (1986). Teacher Education and Teacher Training: The Rush to Mandate. *Phi Delta Kappan 67*: 447–451.

Armor, D. (1972). School and Family Effects on Black and White Achievement: A Re-examination of the USOE Data. In F. Mosteller and D. P. Moynihan (Eds.), *On Equality of Educational Opportunity.* New York: Random House.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A. (1976, August). In E. Pauly and G. Zellman (Eds.) *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Report prepared by RAND for the Los Angeles Unified School District. Report Number R–2007–LAUSD.

Ballou, D., and Podgursky, M. (1997). *Teacher Pay and Teacher Quality*. Kalamazoo, MI: The W. E. Upjohn Institute for Employment Research.

Brophy, J. (1986). Teacher Influences on Student Achievement. *American Psychologist 41*(10): 1069–1077.

Cizek, G. J. (1995, spring). On the Limited Presence of African American Teachers: An Assessment of Research, Synthesis, and Policy Implications. *Review of Educational Research 65*(1): 78–92.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A.M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Darling-Hammond, L. (1997). *Doing What Matters Most: Investing in Quality Teaching.* Kutztown, PA: National Commission on Teaching and America's Future.

Darling-Hammond, L., and Hudson, L. (1989). Teachers and Teaching. In R. J. Shavelson, L. M. McDonnell, and J. Oakes (Eds.), *Indicators for Monitoring Mathematics and Science Education.* Santa Monica, CA: RAND.

Doyle, W. (1986). Classroom Organization and Management. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*, 3rd Edition. New York: MacMillan.

Ferguson, R. F. (1991, Summer). Paying for Public Education: New Evidence on How and Why Money Matters. *Harvard Journal on Legislation 28*(2): 465–498.

Ferguson, R. F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap*. Washington DC: Brookings Institution Press.

Ferguson, R. F., and Ladd, H. F. (1996). How and Why Money Matters: An Analysis of Alabama Schools. In H. F. Ladd (Ed.), *Holding Schools Accountable: Performance Based Reform in Education*. Washington, DC: Brookings Institution Press.

Gifford, B. R. (1985). Teacher Competency Testing and Its Effect on Minorities: Reflection and Recommendations. In E. E. Freeman (Ed.), *Educational Standards, Testing and Access: Proceedings of the 1984 ETS Forty-Fifth Invitational Conference.* Princeton, NJ: Educational Testing Service.

Greenwald, R., Hedges, L. V., and Laine, R. D. (1996). The Effect of School Resources on Student Achievement. *Review of Educational Research 66*: 361–396.

Hanushek, E. A. (1986, September). The Economics of Schooling: Production Efficiency in Public Schools. *Journal of Economic Literature 24*: 1141–1177.

Hanushek, E. A. (1992). The Trade-Off between Child Quantity and Quality. *Journal of Political Economy 100*: 84–117.

Hanushek, E. A. (1996). A More Complete Picture of School Resource Policies. *Review of Educational Research 66*: 397–409.

Irvine, J. (1990). *Black Students and School Failure*. Westport, CN: Greenwood.

Jencks, C., and Phillips, M. (Eds.) (1998). *The Black-White Test Score Gap*. Washington DC: Brookings Institution Press.

Kain, J. (1995). *Impact of Minority Suburbanization on the School Attendance and Achievement of Minority Children.* Final Report of Work Completed Under Spencer Foundation Small Grant. Harvard University, Department of Economics.

Kain, J. (1998, October). *The Impact of Individual Teachers and Peers on Individual Student Achievement.* Working paper. Cecil and Ida Green Center for the Study of Science and Society, University of Texas at Dallas.

Kain, J., and Singleton, K. (1996, May/June). Equality of Educational Opportunity Revisited. *New England Economic Review,* pp. 87–111.

King, S. H. (1993, summer). The Limited Presence of African-American Teachers. *Review of Educational Research 63*(2): 115–149.

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., and Olsen, R. J. (1991). *Who Will Teach? Policies That Matter*. Cambridge, MA: Harvard University Press.

Porter, A.C., and Brophy, J. (1988, May). Synthesis of Research on Good Teaching: Insights from the Work of the Institute for Research on Teaching. *Educational Leadership,* pp. 74–85.

Pressman, H., and Gartner, A. (1986, Summer). The New Racism in Education. *Social Policy 17*(1): 11–15.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (1998). *Teachers, Schools and Academic Achievement.* Paper presented at the January 1998 Annual Meeting of the American Economic Association, Chicago, IL.

Sanders, W. L., and Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.* Research Report. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1998). The Tennessee Value-Added Assessment System (TVAAS): A Quantitative, Outcomes-Based Approach to Educational Assessment. In J. Milliman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, CA: Corwin Press.

Shepard, L. A., and Kreitzer, A. E. (1987, August-September). The Texas Teacher Test. *Educational Researcher*, pp. 22–31.

U.S. Department of Education. (1996). *Digest of Education Statistics*. National Center for Education Statistics.

Response: Two Studies of Academic Achievement

Robert M. Hauser
University of Wisconsin-Madison

The papers by Meredith Phillips (1998) and by Ronald F. Ferguson with Jordana Brown (1998) exemplify the best of contemporary educational policy research.[1] First, they focus on important questions: What are the sources of differentials in academic achievement between racial-ethnic groups in the United States? When do these differentials appear in the course of children's development? What is the role of family and school factors in the development of these differences? How can we best measure, understand, and reduce the differentials? How, if at all, do teacher qualification test scores—or other test scores—affect student learning? Should such test scores be used as a threshold for entry into the teaching profession? What are the effects of such tests on the qualifications of new entrants to teaching and on differentials in the test scores of teachers from majority and minority groups? Will smaller differences between the qualification test scores of majority and minority teachers lead to smaller differences in student achievement? What are the advantages and disadvantages of alternative measures of teacher quality?

Second, both papers use a wide array of evidence. Phillips focuses on new data from the Prospects study, but she—along with her collaborators in related work—actually draws on much of the accumulated evidence of trends and differentials in student achievement in the United States. Ferguson and Brown focus primarily on an important body of data on teacher test scores and student achievement for school districts in Texas, but they also draw on data from other states—notably Alabama—and from other recent studies of teacher qualifications and student performance. One need only think back to the mid-1960s, when the "Coleman-Campbell report" (1966) provided the *only* national

[1] The opinions expressed in this paper are those of the author. Address comments to Professor Robert M. Hauser, Department of Sociology, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706.

data on educational resources and academic achievement, to realize that we have come some distance.

Third, both papers are methodologically and statistically sophisticated, with the authors arranging and examining evidence in new ways and, at the same time, not letting their work become model-driven to the point where they completely lost sight of the data or of the limits of their data in addressing their central questions. Indeed, both Phillips and Ferguson and Brown focus as much on better ways to ask their questions as on the important findings of their research.

Enough of generalities, what about the papers?

Response to the Phillips Paper

Phillips makes six main points. They are worth repeating, although I will quibble a bit with some of them.

The first point is as follows: “Traditional socioeconomic factors do not overlap with ethnicity as much as many people assume.” Ethnic differences in achievement are not easily reducible to socioeconomic or other social differences in academic achievement. Phillips observes that the reductionist view has been sustained in part by the political sensitivity of black-white differences. Thus, many researchers have tried to explain the gaps, as Phillips notes, by black-white differences in levels of family advantage, neighborhood poverty, or urban-suburban location. But these factors do not account for the test score gap.[2]

The second point is, “We should focus our surveys mainly on elementary school students rather than on high school students.” I think this is a bit overdrawn. To the degree that our focus is on academic achievement, the available evidence points to the malleability of learning in the early years. That is important. But we ought not to forget adolescence—recall the success of recent years in changing course content and requirements in high school—as well as the

[2] At the extreme, I have seen one leading economic scholar argue against adjusted statistical comparisons of educational outcomes between blacks and whites on the ground that there is not sufficient overlap of socioeconomic background to justify this form of comparison—a proposition that is patently contradicted by the evidence of overlap between distributions of social and economic standing in the black and white populations.

wider array of outcomes that determine what happens to youth when they leave high school (Hauser 1991). The downward drift of starting points of the major national longitudinal studies—from NLS 1972 to HS&B in the 1980s and NELS in the 1990s—has been a beneficial evolution. But we ought not to lose such samples as they age, no matter how young they are when we start. I will come back to this point again in discussing Phillips' fifth point.

The third point is as follows: "We should test children in both the fall and the spring of each school year." As Phillips notes, her evidence on this point from the Prospects study is compelling, and it builds on a decades-old history of similar findings. Why has this source of black-white test score differences not become a focus of public policy? What would it take to accomplish that? Need we wait until achievement test scores sink so low that the public approves test-based grade retention on a massive scale before we put any real money into summer school?

But I would question the calculus of Phillips' statistical comparisons of learning in summer school and during the school year. Such comparisons read *as if* score gains during the school year are the work of schools alone, while summer gains or losses are the work of families alone. Consider alternative assumptions: Suppose learning is linear in exposure to learning environments. Students do not leave their families during the academic year; they spend more time in school and somewhat less with families. Suppose we ignore summer school, and attribute summer gains or losses to families. Then, summer changes reflect the effects of families (including peers and neighbors), while changes during the academic year reflect the combined effects of schools and families. Assume, further, that exposure to school and family is equal throughout the academic year. Now, for example, look at the top row of table 6. Three months of family-only exposure in the summer produces a black loss of 20.47 points. This implies a loss of 20.47/6 = 3.41 points per month during the school year, assuming summer is 3 months long and family exposure is half as great during the academic year as in the summer. The implied loss is 3.41 x 9 = 30.71 points during the academic year. Since black children gain relative to whites during the academic year—by 9.27 points—the implication is that the annual effect of schooling is 9.27 + 30.71 = 39.98 points. In this account, schooling plays an enormously effective role in reducing black-white test score differences.

Please do not take this account too seriously. In particular, the test used in the Prospects study is vertically equated to show larger gains at low than at high performance levels. Thus, the score gains of African American students are not strictly comparable to those of majority students. My assumptions and calculations are no more than illustrative. But they are, I think, worth thinking about. What is the role of schools in learning relative to families during the school year? During the summer? How could we learn more about it? Do family effects really offset school effects, or are they complementary? If so, how do we explain the summer deficits? What would be the long-term benefits of year-round schooling, and how could we realize them?

Phillips' fourth point is, "Tests of seemingly similar skills … sometimes yield very different estimates of ethnic differences in achievement." Here, the evidence provided by Phillips (in table 7) appears supportive, but I am not sure that it is strong enough. The problem is that the measures of math concepts and math computations are not independent, so simple comparisons of means and their reported errors are not appropriate to test differences in the effects of ethnicity on the outcomes. A bit more modeling is required.

The fifth point cited is as follows: "Different surveys of apparently similar populations sometimes yield contradictory results." I am not at all convinced by the comparison of children of the NLSY with those of Prospects in table 8. The key issue here is "apparently similar populations." The CNLSY is a household-based survey, and children of women in the NLSY of 1979 have passed through school over a period of years, assuming that Phillips has captured the experience of those children in full. Those children do not represent all children in the birth cohorts because children in the same years may be born to mothers outside the cohorts of the NLSY. Children of the NLSY are subject to attrition from both the parent and child samples. Children of the NLSY do not include children of recent immigrants from the same or different cohorts as the mothers of the NLSY. I am not at all sure that it is worth trying to reconcile all of the differences between Prospects and CNLSY; I am reasonably sure that the fact that the surveys yield discrepant findings does not in itself justify a call for multiple, independent survey operations.

The sixth point is, "The vocabulary gap between African Americans and European Americans is already large by the time children are three years old." I agree, and that's why we are here.

What survey research designs might address Phillips' concerns? I would suggest, and not for the first time, that the need for replicate observations and for alternative methods should be met by the regular initiation of new, and perhaps modestly sized, longitudinal cohort surveys—and not by larger, one-time-only or once-per-decade surveys. I have made the same proposal for studies of adolescent development. We ought to be initiating cohort surveys close to birth every year—or every other year—as a means of improving our "who, what, when" understanding. Such surveys should be stratified by ethnic origin, differentially sampled. And they should provide opportunity for experimentation with alternative test (and questionnaire) content and observational designs, as well as opportunity for core content stable enough to permit aggregation of findings across cohorts to yield greater statistical power. There is already a considerable literature on the need for such surveys and on possible designs (National Research Council 1995). We need not reinvent it here.

Response to the Ferguson and Brown Paper

Ferguson and Brown (1998) focus primarily on the effects of teacher test scores on achievement test scores in Texas. They briefly consider other measures of teacher effectiveness: classroom observation—which they dismiss as too costly and of doubtful validity—and direct observation of student gains in test scores. They dismiss the latter as requiring years of observed data but note, "When the data are adequate and the analysis is done appropriately, [this is] probably the best that we can do at measuring the effectiveness of individual teachers." I agree about the validity of this method and wonder why it is not viewed as more practical for the evaluation of teachers beyond point-of-hire. Many of us—as college and university academics—have had judgments made about our effectiveness and competence on the basis of accumulated dossiers. To be sure, these are lists of books, papers, and talks, rather than raw scores, but the principle is the same.

Ferguson and Brown's analyses of the Texas data are for school districts as units. Since the ratio of student test score standard deviations at the district level to those at the individual level is as 1 to 3—and similarly for teacher test scores—much of the analysis of statistical findings passes transparently from one level to the other. This is convenient, but perhaps too much so. Districts are not students, and the specification problems that seem obvious to us when we think about determinants of individual test scores—many of which have been satisfied by Ferguson and Brown—may not be the right ones to solve at their level of aggregation. They know it, and they say it, but it remains a puzzle. The 1 to 3 ratio of standard deviations has an important implication that goes unstated in the text: 90 percent of the variance in student test scores, like 90 percent of the variance in teacher test scores, lies within districts. There is a lot of room for specification error at the latter level to escape our notice. Think, for example, of the issue of validity of certification tests such as the TECAT, and ask yourself whether the Texas school district data bear on that issue.

I also worry about the fact that the data are from 1985–86, just at the time teacher testing was introduced. What has happened to the distributions of teacher test scores, both within and between districts, since that time? Are the standard deviations of 1985–86 the right metric for us to use in thinking about policies in the late 1990s?

Ferguson and Brown are both clever and wise in their statistical analysis. I particularly commend the use of a “difference of differences” estimator of the effect of teacher test scores on student gains, reported in table 1. Similarly, I like the fact that they help us look directly at the data in figure 1, which is one of the most fascinating statistical graphics I have seen in some time. It is a striking example of what demographers would call a synthetic cohort analysis, in which variation of test scores across grade levels within a single year is taken as a proxy for variation in achievement within a single cohort across its progression through grade levels. What the figure appears to show is that academic achievement tends to become consistent with initial teacher scores in the cases where substantial inconsistencies occur in the lower grades. There are some reasons to be wary of this finding: The number of inconsistent districts is very small, and we do not know why they are inconsistent. And the data pertain to synthetic cohorts, not real changes in academic achievement across time.

Finally, if I understand the graphic correctly, this is one case where the distinction between measurements in standard deviations at the district versus individual levels really makes a difference. Imagine rescaling figure 1 in standard deviation units of individual test scores. In this case, if I follow the arithmetic, a consistent, 10-year improvement of teacher scores by two standard deviations—how feasible is such a gain?—would accumulate to 0.57 standard deviations. That is a substantial fraction of initial black-white differences in test scores. Is the evidence strong enough to support such a conclusion?

I would add that this is a striking, but perhaps too limited, example of our need to get closer to the data. We should be doing a great deal more exploratory data analysis, even in situations where we think we know how to model data successfully. Even in large, longitudinal surveys, that may help us as much in putting together a coherent story as any number of smaller, even smaller and richer, studies. Model the data, but also look at the data.

One other question about the Ferguson-Brown paper strikes me as particularly important. It is mentioned at the close of the paper. How much of the measurable difference in teacher effectiveness can be attributed to test score differences? That is, suppose we ran the dummy variable regressions of student test score change on “teacher” as described at the beginning of the paper. What would happen to the coefficients of teachers as their test scores enter the equation? And what other teacher characteristics would explain the remaining effects? We might ask, also, whether the effects of teachers’ test scores are diagnostic or causal. That is, do they truly account for teachers’ effectiveness, or are they merely sound evidence to be used in screening potential teachers? One way or the other, what are the costs and benefits of improved supervision and training relative to—or complementary to—the skills and knowledge that teachers initially bring to the job? For example, what should we make of the evidence that the support of teaching and teachers is a major impediment to the success of standards-based reform?

Persistent Issues in Educational Policy Research

These two fine papers also remind me of potential weaknesses and points of contention in contemporary educational policy research. Two of these points of contention are the centrality of test scores as educational outcomes and a possible failure to respect the limits of observational data in answering policy questions that can only be answered in the language of cause and effect. This is not news, but I think the main points bear repeating.

We are here to think about academic achievement—how it is produced, how it becomes differentiated, how to measure it, how to measure its production. This is all well and good: I do not want to be one of those miserable critics who say that we should not be here doing what we are doing today. Student learning is a main objective of schooling, and achievement tests are a great social invention as well as our main way of measuring student learning. But we in the research and policy community can focus too much on tests and testing. The focus on student learning as a key outcome of schooling devolves into a focus on student test scores as a key outcome of schooling, and the latter may devolve into a focus on student test scores as the *only* outcome of schooling. To paraphrase Vince Lombardi's remark concerning winning, "Test scores are not the main thing; they are the only thing." As researchers, we may learn all about academic achievement—and little else. As a nation, we may get what we wish for and live to regret it.[3]

If we know all too little about educational production functions, we should be even more humble about our understanding of what makes people healthy, wealthy, and wise. For 30 years I have been watching as the 10,000 students in the Wisconsin Longitudinal Study have marched through life. This is the same cohort of 1957 high school graduates portrayed in the situation comedy, *Happy Days*. I have learned two things as I (along with my colleagues) have watched trajectories of schooling, jobs, and family lives, and of states of depression and well-being and of health and disease. The first—to use a quip by Paul Siegel from some years ago—is that everything that happens to you before your sixteenth birthday affects everything that happens to you after your sixteenth birthday by way of the amount of schooling that you finish. The second is that adolescent test scores provide no exception to the rule.[4] Education is not just test scores, and we should not wish to make it so. Education is a fascinating bundle of learning and motivation, of values and skills, of behaviors and—yes—certification, and the easy part of our job is to unbundle it. The hard part is not to lose sight of the whole. In a smaller, but older longitudinal study, the late social psychologist, John Clausen (1991, 1993) summarized the key to the good life as *planful competence*—a combination of academic success with responsibility and motivation.

All of this broadens the subject without any reference to the demography of schooling, with which the connections with academic achievement are pervasive, complex—and largely ignored. To go back to the problem of getting what we wish for, I think it is fair to say that we in the research policy community have aroused, and now bear responsibility for moderating, the national mania for achievement testing. Read *High Stakes: Testing for Tracking, Pro-*

[3] I am reminded of the urban renewal program of the late 1950s and early 1960s, in which the goal was "decent, safe, and sanitary housing." That is just what we got, but only for a short time, and what we did not get was healthy, viable communities.

[4] For example, see Hauser and Sweeney (1997).

motion, and Graduation, the report of the National Research Council (1999), if you want to learn more about both of these last two points.

In the closing passage of her paper, Meredith Phillips rightly observes the distinction between *understanding* the growth of differentials in academic achievement—the main focus of her work—and *changing* those differentials—a task for which, she argues, we should turn to large-scale field experiments. Similarly, Ferguson and Brown muse about the limits of econometric methodology in explicating the role of teacher qualifications in student achievement. I would put the matter somewhat differently, i.e., observational studies, even those designed and carried out to the highest standards, are mainly useful in telling us what has happened, when, and to whom. I am all in favor of putting such accounts into the form of statistical models, to the extent justified by the data and by prior knowledge and plausible assumption. Such exercises are most valuable—witness Meredith Phillips' compelling finding that summer deficits dominate winter surpluses of learning among black schoolchildren. But they do not tell us "how to fix it." The language of causality provides a useful way of thinking about the world, but we ought not to invest it with more belief than our research designs and evidence can sustain.[5]

[5] On the other hand, research on experimental and nonexperimental methods of evaluating welfare policies and reform provides equally cautionary evidence about the value of observational data—especially when we take the final leap between theory and practice.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A.M., Weinfeld, F. D., and York, R. L. (1966) *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Clausen, J.A. (1991). Adolescent Competence and the Shaping of the Life Course. *American Journal of Sociology 96*(4): 805–842.

Clausen, J.A. (1993). *American Lives: Looking Back at the Children of the Great Depression.* New York: The Free Press.

Hauser, R.M. (1991, fall-winter). What Happens to Youth After High School? *IRP FOCUS 13:* 1–13.

Hauser, R.M., and Sweeney, M.M. (1997). Does Adolescent Poverty Affect the Life Chances of High School Graduates? In G. Duncan and J. Brooks-Gunn (Eds.), *Growing Up Poor* (pp. 541–595). New York: Russell Sage Foundation.

Herrnstein, R.J., and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: The Free Press.

National Research Council. (1995). *Integrating Federal Statistics on Children.* Washington, DC: National Academy Press.

National Research Council. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation.* In J.P. Heubert and R.M. Hauser (Eds.), Washington, DC: National Academy Press.

Section III.
Relating Family and Schooling Characteristics to Academic Achievement

Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88..... 169
Dominic J. Brewer and Dan D. Goldhaber

School-level Correlates of Reading and Mathematics Achievement in Public Schools 189
Donald McLaughlin and Gili Drori

Response: Opportunities for Design Changes 237
Valerie E. Lee

Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88

Dominic J. Brewer and Dan D. Goldhaber
The Urban Institute

Introduction

The mission of the National Center for Education Statistics (NCES) is to collect "…statistics and information … in order to promote and accelerate the improvement of American education." To help achieve this, over the past 20 years NCES, as well as its predecessors, has greatly expanded its collection of longitudinal data. As a result, researchers have gained a better understanding of educational practices and the underlying complex relationships between students, schools, and teachers. In the absence of large-scale, randomized experiments to determine the effectiveness of educational interventions and resources, analyses of nonexperimental data can provide insights that help policymakers allocate scarce resources and enable practitioners to improve student achievement.

Analyses of NCES data have generated a large amount of literature focusing on the key determinants of student achievement and the effects of programs and policies. Educational productivity studies have focused on the overall effects of spending on schools and on the effectiveness of particular educational inputs (Monk 1992). In particular, these studies examine how per pupil expenditures and school, teacher, and class characteristics (e.g., school demographics, teacher degree levels, and class size) affect student outcomes such as test scores. This research has spawned what is commonly referred to as the "does money matter?" debate. Much of this controversy has been shaped by older studies cited in Hanushek (1986) that rely on cross-sectional and aggregated data. Although the issue is not yet settled, the availability of longitudinal data, as well as the use of more sophisticated statistical methods, has advanced our knowledge in the area. For example, earlier studies examining the impact of class size on student outcomes used data aggregated to the school and school district levels. However, recent longitudinal data are specific enough to enable researchers to use actual class sizes rather than school or district average pupil/teacher ratios. Detailed longitudinal data have also allowed scholars to examine controversial educational reforms such as tracking and school choice.

In this paper commissioned by NCES,[1] we were asked to illustrate how we have used recent NCES longitudinal databases in our own research to investigate issues of "educational productivity," broadly defined as the relationship between school resources and educational practices and student achievement.[2] We argue that the *National Educational Longitudinal Study of 1988* (NELS:88) represents a substantive improvement over previous longitudinal data collection efforts for research and policy purposes. These enhancements permit the estimation of a wider variety of statistical models of the determinants of student achievement and allow researchers to test important hypotheses about educational practices that have implications for policy. In particular, the ability to link individual students to detailed background information about their teachers proves critical to interpreting the results of standard education production functions. As an example, our own research demonstrates (Brewer and Goldhaber 1996; Goldhaber and Brewer 1997a, 1997b) that subject-specific teacher background in mathematics and science is systematically related to student achievement in these subjects, even though teachers' higher degrees in general are not. We suggest a number of further improvements to future NCES longitudinal data studies including collection of more refined information on teacher characteristics and ability, increased frequency of follow-ups, and more student-level observations per teacher.

The paper begins with a brief introduction to the education productivity literature, followed by a discussion of some of the advantages of NELS:88 over previous NCES data, focusing on the substantive findings on teacher subject-specific preparation. Next, we argue that these data have permitted the

[1] This paper is based primarily on the author's research on teacher qualifications previously published in the *Journal of Human Resources, Advances in Educational Productivity, Developments in School Finance 1996,* and *Education Economics.*

[2] Hence we make no attempt to be comprehensive in reviewing other research here. That is not the purpose of this paper.

estimation of a broader set of statistical models, including hierarchical linear models, fixed and random effect models, and models with selectivity. The conclusion makes a number of suggestions for future data collection.

School Resources and Student Achievement

The ultimate reason to collect data is to influence public policy in a positive way. Thus, researchers are interested, among other things, in furnishing policymakers with the information required to make prudent resource allocation decisions and to understand which educational interventions work. Dating back to the 1966 “Coleman report” (Coleman et al.1966), there have been numerous studies on how investments in educational resources affect student performance and labor market outcomes. This line of research falls under the broad heading of “educational production functions.”[3] Most educational production function studies seek to explain variance in standardized test scores at student, school, or school district levels by estimating multiple regression models that regress student outcomes on individual and family background variables and school inputs. The broad conclusion of this body of work is that individual and family traits explain the vast majority of variance in student test scores and that schools play a lesser role. Eric Hanushek notes that these studies as a whole show that “differences in [school] quality do not seem to reflect variations in expenditures, class sizes, or other commonly measured attributes of schools and teachers” (Hanushek 1986, 1142).

He concludes that there is “no strong evidence that teacher-student ratios, teacher education, or teacher experience have an expected positive effect on student achievement” and that “there appears to be no strong or systematic relationship between school expenditures and student performance” (Hanushek

[3] The notion that there is an estimable education production function for a set of individuals within or across classes or teachers or schools or school districts is not unchallenged (Monk 1992). Like any model, the education production function is certainly a simplification of reality, but it is a useful tool. This is particularly true for policy purposes because most applications focus on manipulable, measurable inputs rather than on intangible variables or amorphous constructs like “school climate” that are difficult to translate into practical recommendations.

1986, 1162).[4] Hanushek's interpretation of the literature suggests that (public) schools have a suboptimal allocation of resources (allocative inefficiency), that they do not operate on the production possibility frontier (technical inefficiency), or both. In these cases additional teacher inputs or smaller class sizes would not necessarily imply higher output, *ceteris paribus*. This result does *not* imply that schooling resources *never* affect student achievement positively, simply that, given the way public schools are organized, additional resources do not make much systematic difference.

The view that observable school inputs, and teachers in particular, do not positively impact student achievement rests on somewhat shaky empirical ground. Hanushek's conclusion is based primarily on older work, and there are good reasons to believe that much educational productivity research completed in the 1970s had major deficiencies. One problem is likely to be that key variables may have been omitted from estimated test score models, potentially leading to biased coefficient estimates of the included variables.[5] Missing information and crude proxies for many schooling inputs in older data make this likely. For example, many early studies were unable to control for prior achievement using a "pre-test" score to net out individual ability (Boardman and Murnane 1979; Hanushek 1979; Hedges, Laine, and Greenwald 1994).

Additionally, schooling inputs—notably class size and expenditures per pupil—are measured with some degree of error. This error arises, in part, from aggregation of variables to the school or district level (Hanushek, Rivkin, and Taylor 1996). For instance, rather than class size, studies often utilize total school enrollment divided by the total number of teachers (or professionals) as

[4] Hanushek concluded that there is no *systematic* relationship between observable schooling resources and student test scores, at least at current levels of resource utilization, by noting the direction of estimated input effects (teacher/pupil ratio, teacher education, teacher experience, teacher salary, per pupil expenditures, administrative inputs, and facilities) on student achievement, along with whether they were statistically significant, and simply tallying ("vote counting") the number of statistically significant positive and negative coefficients. A "meta-analysis" by Hedges, Laine, and Greenwald (1994) using the same set of studies reviewed by Hanushek reached a very different conclusion. Their basic argument was that the *pattern* of estimated coefficients in these studies suggested there were indeed systematic positive effects, although it is not clear that the alternative interpretation gives any clearer guidance for policymakers.

[5] Omitted variables and aggregation bias problems in the context of educational production functions are discussed in a more formal fashion in Goldhaber and Brewer (1997b); see also Hanushek (1979) and Hanushek, Rivkin, and Taylor (1996).

an average pupil/teacher ratio. This is problematic given that there is considerable variation in class size *within* schools as well as *between* schools. Variables representing school and teacher "quality" used in most production function studies are typically very crude. For example, teacher degree level and years of experience may be only weakly related to teaching skill. Degree alone does not distinguish among diplomas given at high- and low-quality colleges or when the degree was granted, nor does it convey any information about college major, certification requirements fulfilled, or subsequent professional development. Teacher motivation, enthusiasm, and skill at presenting class material influence students' achievement, but are difficult traits to accurately measure and are, thus, omitted from standard regression analyses.

Recent Longitudinal Data: Reducing Omitted Variables and Aggregation Bias

Advances in statistical techniques and the collection of two large-scale longitudinal databases by NCES, *High School and Beyond* (HS&B) and the *National Educational Longitudinal Study of 1988* (NELS:88), have allowed researchers to learn more about how school resources impact students. In this section we first review the advances in these data and introduce our work on teacher subject-matter preparation, which reduces the omitted variables and aggregation bias problems inherent in earlier studies.

HS&B and NELS:88

High School and Beyond (HS&B) was one of the first large-scale longitudinal databases. A cohort of students were tested in both tenth and twelfth grades, permitting researchers to use a "value-added" methodology—examining how much students learned between two points in time (as measured by standardized tests), using a pre- and post-test. Unfortunately, teacher data were only collected in 1984, two years after the students had graduated and, therefore, could not be linked back to particular students. Hence, like previous work, statistical models estimated using individual level test scores had to rely on school-level measures of variables such as class size and teacher experience (Ehrenberg and Brewer 1994).

The NELS:88 data represent a substantial improvement over HS&B because they include concurrent, detailed school and teacher data collected in such a way that researchers can link students to their particular teachers and classes. The NELS:88 study is a nationally representative survey of about 24,000 eighth grade students conducted in the spring of 1988, with follow-ups conducted in 1990, 1992, and 1994. At the time of each survey, students took one or more subject-based tests in math, science, English, or history. The tests were carefully designed to avoid “floor” and “ceiling” testing effects and were put on a common scale using Item Response Theory.[6] Linked student-teacher-class data allow an investigation of the impacts of specific class size, teacher characteristics, and peer effects on student achievement. NELS:88 is constructed in such a way that students can be linked to data gathered in a separate teacher survey that provides information on the teacher’s background and teaching methods and curricula used in the particular class the tested student is taking. This represents a major advance.

Of course, there are still some important deficiencies in the NELS:88 design. For example, the study skips a year in sampling students so information on the characteristics of schools, teachers, and classes is missing for the ninth and the eleventh grades. Because information on the ninth and eleventh grades is left out of the survey, studies that use the NELS:88 may suffer from omitted variables bias. The direction and magnitude of this bias depends on the relationship between the tenth and/or twelfth grade characteristics included in the survey and those in the excluded grades. Further, in each follow-up to the eighth grade base year, there are fewer students per class (i.e., there is a fanning out of the original set of eighth grade students to multiple teachers and classes by the tenth grade). As a result, in some cases there are as few as one student per class and teacher in the follow-up surveys. This makes it more difficult to distinguish between teacher and class effects.[7]

Several researchers have taken advantage of the ability to link students to their classes and teachers. For example, the magnitude of the effects of class size on student achievement has long been debated, but there is currently widespread interest in class size-reduction policies at federal and state levels (Parrish and Brewer 1998). In the absence of large-scale experimental data, such as

[6] For more information on this methodology, see Rock and Pollock (1991).

[7] This is not a problem for some kinds of statistical model (e.g., ordinary least squares, random effect models). In all of our work we use tenth grade school, teacher, and class variables, and in most cases impose no restrictions on the minimum number of student observations per class or per teacher.

Project STAR in Tennessee, nonexperimental production function type studies will continue to be important, so it is necessary to have data that allow for the best possible test of the relationship between class size and student outcomes.

Many of the class size studies cited by Hanushek (1986) have found no relationship between class size and student achievement. Some have even found that achievement is higher in larger classes—this is clearly counterintuitive. Using NELS:88, Akerhielm (1995) shows that students are not randomly distributed across classes within schools. Low-achieving students tend to be assigned to smaller classes. Without accounting for this nonrandom assignment of students to classes, there appears to be a positive relationship between class size and achievement. However, when statistical models incorporate this nonrandom assignment, the relationship becomes negative (i.e., smaller classes result in higher achievement) between class size and student achievement. Because NELS:88 links students to a particular class, Akerhielm was able to measure the actual class size rather than an aggregate pupil/teacher ratio.

Teacher Subject-specific Preparation

The ability to link students to their particular teachers afforded by NELS:88 has also allowed researchers to better understand how teachers affect students. Ehrenberg, Goldhaber, and Brewer (1995) were able to shed some light on the issue of “role models” in education by investigating whether the race-ethnicity and gender of teachers impact student test scores.[8] Similarly, NELS:88 permits an investigation of whether the subject-specific preparation of teachers affects student achievement. The National Commission on Teaching and America’s Future in 1996 reported that one-fourth of high school teachers lack college training in their primary classroom subject. Underlying this concern about out-of-field teaching is the assumption that teachers with degrees in the subject that they teach are more effective. Although this may seem a commonsense proposition, there is relatively little quantitative work on the relationship between educational outcomes and teacher subject-specific

[8] This research shows that, on balance, teachers’ race-ethnicity and gender are more likely to influence teachers’ subjective evaluations of their students than to affect student achievement as measured by standardized tests. The important point here, however, is that this research would not be possible without teacher data directly tied to individual students.

preparation, because most data used for these analyses do not contain this information.[9]

Using samples of students in four subjects,[10] Goldhaber and Brewer (1997a) estimated standard education production function models in which a student's tenth grade test score in a subject is regressed on (1) individual and family background variables (e.g., sex, race-ethnicity, parental education, family structure, family income, and eighth grade test score); (2) tenth grade school level variables (e.g., urbanicity, regional dummies, school size, the percentage of students who are white, the percentage of students from single-parent families, and the percentage of teachers with at least a master's degree); (3) tenth grade teacher variables (e.g., sex, race-ethnicity, years of experience at the secondary level, whether the teacher is certified, and his or her degree level); (4) and tenth grade class-level variables (e.g., class size and percentage of minority students in the class). The results from these models demonstrate two important things. First, school level aggregates, such as the percentage of teachers in a school with at least a master's degree—the extent of information on teachers available in previous data, are statistically insignificant in all estimated statistical models regardless of subject.

Second, the ability to include subject-specific teacher degree and certification information is critical to interpreting the results of these statistical models. This is illustrated in table 1. Columns (1), (3), (5), and (7) of the table show the estimated regression coefficients of the teacher variables included in the model, while columns (2), (4), (6), and (8) show the results when we include more refined subject-specific teacher characteristics (whether the teacher is certified in his or her subject area and whether the teacher has BA or MA degrees in his or her subject area). These variables allow us to distinguish between teachers who have BA or MA majors in the subject they teach, those who have certification in the subjects they teach, and those who do not have subject-specific training. In the models reported in columns (1), (3), (5) and (7), *years of teaching experience* is not a statistically significant item, nor is whether the teacher

[9] Monk and King (1994) report that teacher subject matter preparation in mathematics and science does have some positive impact on student achievement in those subjects. Again, this insight is made possible only because the *Longitudinal Survey of American Youth* (LSAY) data that they used links students with individual teachers.

[10] The sample sizes for the four subjects were 5,113 students in math; 4,357 students in science; 6,196 students in English; and 2,943 students in history.

Table 1. Comparison of Selected Coefficients from Educational Production Functions[a] (absolute value of *t*-statistic)

	Math		Science		English		History	
Teacher Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of experience at secondary level	0.018 (1.5)	0.013 (1.1)	0.007 (0.7)	0.007 (0.6)	-0.007 (0.6)	-0.007 (0.7)	0.025 (1.6)	0.025 (1.7)
Certified	-0.511 (0.9)	-2.343 (2.3)	0.140 (0.3)	-0.827 (1.2)	-1.267 (1.9)	-0.645 (0.7)	0.170 (0.2)	0.142 (0.1)
Certified in subject	-	2.172 (2.2)	-	1.130 (1.2)	-	-0.685 (0.9)	-	0.035 (0.0)
BA or more in subject	-	0.769 (3.6)	-	0.683 (3.3)	-	0.130 (0.3)	-	-0.243 (0.8)
MA or more degree	0.247 (1.2)	0.052 (0.2)	0.030 (0.2)	0.023 (0.1)	-0.070 (0.4)	-0.085 (0.4)	-0.038 (0.1)	-0.056 (0.2)
MA or more in subject	-	0.595 (2.1)	-	0.002 (0.0)	-	0.078 (0.3)	-	0.101 (0.3)
Sample Size	5113	5113	4357	4357	6196	6196	2943	2943
Adjusted R^2	0.766	0.767	0.377	0.378	0.605	0.605	0.275	0.274

[a] Models also include the following variables: sex, race-ethnicity, parental education, family structure, family income, eighth grade test scores, urbanicity, regional dummies, school size, the percentage of students who are white, the percentage of students from single-parent families, the percentage of teachers with at least a master's degree, tenth grade teacher sex and race-ethnicity, and tenth grade class size and percentage of minority students in the class.

NOTE: With these sample sizes, any *t* statistic greater than 1.645 indicates the estimate is statistically significant at the 10 percent level, and 1.960 indicates the estimate is statistically significant at the 5 percent level.

Reproduced from Goldhaber and Brewer (1997a).

has a master's degree—implying that teachers with master's degrees are no more (or less) effective than those without advanced degrees. The results for teacher certification are similar in that we find the coefficient on teacher certification to be statistically insignificant (except in English, where teacher certification is significant and negative). If these were the only variables available, one might erroneously conclude that teacher preparation does not matter.

By contrast, in math and science, teacher subject-specific training has a statistically significant impact on student test scores in those subjects [columns (2) and (4)]. A teacher with a BA or an MA in math has a statistically significant *positive* impact on students' achievement relative to teachers with no advanced degrees or degrees in non-math subjects.[11] Further, these findings appear to reflect subject-specific training rather than simply teacher ability.[12]

Table 2 shows the estimated effects of model specification on predicted tenth grade achievement scores in math and science (we do not show English and history, because none of the subject-specific variables were statistically significant). We can infer the magnitude of the effect of teacher training on student achievement by examining the estimated coefficients in the models that include subject-specific information. For example, the effect of a teacher's having an MA in math is the sum of the coefficients of MA and MA major in math. We see the impact of model specification in math and science by comparing columns (1) and (2) for math and columns (3) and (4) for science. In both math and science, a subject-specific BA degree improves student achievement; and the results are even more pronounced for math, where an MA in math also has a statistically significant effect on achievement. In the model with general teacher variables, we predict students (with average characteristics) who have a teacher who is certified in math and has both a BA and an MA in math to have a tenth grade math score of 44.06. However, these same stu-

[11] See Goldhaber and Brewer (1997a, 1997b) for more details on various robustness checks and statistical tests performed on these models. We find no evidence that certification or subject-specific degrees have an effect on student achievement in English or history, where the subject-specific variables were statistically insignificant.

[12] Teacher math and science degrees may serve as proxies for teacher ability. To test this hypothesis, we re-estimated all models, including whether a teacher has a math or science degree in the English and history regressions. If math and science degrees serve as proxies for teacher quality, we would expect the coefficients on these variables to be significant and positive in all of the subject areas, including English and history. This is not the case. Neither the math nor the science degree level variables are statistically significant in the English and history regressions.

Table 2. Effect of Model Specification on Predicted Test Scores[a]

	Math		Science	
	Model with general teacher variables	Model with subject-specific variables	Model with general teacher variables	Model with subject-specific variables
Certification (in subject)	43.94	43.95*	21.79	21.81
BA only (in subject)	43.96	44.21*	21.78	21.99*
MA only (in subject)	44.08	44.57*	21.79	21.78
BA, MA, and certification (in subject)	44.06	44.69*	21.80	22.02

[a] All other variables are measured at their mean value.

* Indicates statistically significant difference from base model with general teacher variables.

Reproduced from Goldhaber and Brewer (1997a).

dents are predicted to have a tenth grade math score of 44.69 when the subject-specific specification of the model is used. The difference between these predicted scores, .63, is about 5 percent of the tenth grade math test standard deviation, a relatively small (but statistically significant) difference. This finding is important, because it suggests that student achievement in technical subjects can be improved by requiring more in-subject teaching. The estimation of statistical models that yield this finding is only made possible by the linked and detailed data available in NELS:88.

Recent Longitudinal Data: Permitting a Broader Class of Statistical Models

In addition to reducing the omitted variables and aggregation biases in earlier production function studies problems (Goldhaber and Brewer 1997b; Hanushek, Rivkin, and Taylor 1996), the NELS:88 data allow researchers to estimate a broader class of statistical models that may have useful implications for understanding educational productivity.

Hierarchical Linear Models

One example of such a model type made possible by NELS:88 is illustrated by Lee and Smith (1997), who investigated the relationship between school size and student achievement using a technique called hierarchical linear modeling (HLM). The basic approach of HLM is to first estimate a within-group model and use the estimated slope coefficients as the dependent variable in a second across-group stage. The HLM technique has become widely used by educational researchers to model effects that are thought to correspond to particular levels or groupings. It may be argued that standard statistical models (e.g., ordinary least squares) yield inefficient estimates of the effects of some schooling variables when those variables affect groups of students jointly. This is because individual level models, which include higher-level effects (such as school climate), may not adequately capture the contextual impact of these higher-level effects. Thus, the NELS:88 data permit these types of models because they contain multiple observations per class, teacher, and school. Several researchers have taken advantage of this data structure to examine different hypotheses.[13]

[13] For instance, see Lee, Smith, and Croninger (1997) for an investigation of how the social and academic organization of high schools affect learning in math and science and Gamoran (1996) for a comparison of public and private school achievement.

Although HLM is intuitively appealing and may yield efficiency gains, it is important to note that, like any other statistical model, it requires a particular data structure and is predicated on a set of assumptions, such as the distribution of the error term. Further, it is only possible to estimate these models when there are multiple observations at each contextual level.[14]

Thus, with the NELS:88 data, HLM utilizes only a sub-sample of all the potential students in the sample and some information is lost. Additionally, the choice over the level of the effect is somewhat arbitrary. For instance, Lee and Smith (1997) specify the effect at the school level but ignore the potential for class-level effects. Similarly, one might argue that the contextual effect is at peer (small group) or neighborhood levels, both of which are typically ignored. Finally, HLM does not permit researchers to handle situations when the first-level outcomes and second-level regressors are jointly determined (endogenous) (Mason 1995). For instance, school quality may play an important role in influencing parental choice of school sector. Unless models explicitly account for this type of selection, they will yield biased coefficient estimates. The bottom line is that although HLM may yield more efficient estimates of contextual effects, it only addresses one of the many problems associated with using nonexperimental data, such as omitted variables bias, sample selection bias, and endogeneity.

Fixed- and Random-Effects Models

The NELS:88 data do allow researchers better opportunities to investigate some of these issues. For example, Goldhaber and Brewer (1997b) use the data to test whether unobservable teacher characteristics cause systematic bias in the estimated effects of observable variables. This is made possible because the structure of the data permits the estimation of fixed- and random-effects models. The results of this work suggest that unobservable teacher characteristics such as motivation and skill are not systematically related to observable teacher characteristics and that this does not cause bias in standard educational production function studies.

In follow-up work, Goldhaber, Brewer, and Anderson (1999) calculate the role of individual, family, school, teacher, and class characteristics in ex-

[14] It is not clear, however, how many observations are required at each level to adequately capture the contextual effects in question.

plaining variance in student (math) test scores. In doing so, they distinguish between the contributions of observable and unobservable factors. They find the vast majority of variance is explained by individual and family background characteristics (about 60 percent). Overall, school, teacher, and class variables, both observed (e.g., class size and teacher certification and degree level) and unobserved (e.g., teacher motivation and parental involvement), account for approximately 21 percent of the variance in student achievement. Of this 21 percent, only about 1 percentage point (or 4.8 percent) is explained by observable educational variables, and the remaining 20 percentage points (or 95.2 percent) are made up of unobservable school, teacher, and class effects.

These unobservable effects may represent variables, such as the climate in the school, the students' peers, the ability of the teacher, or unobservable family background characteristics that are not adequately controlled for in the model.

Selectivity-Corrected Models

The issue of student selection also often arises when researchers attempt to understand how schools impact students. Student selection is another form of omitted variables bias that occurs when individual characteristics not easily quantified in data are associated with a choice made by those individuals. For instance, one educational reform that has recently gained a great deal of attention is the use of educational vouchers (school choice). Proponents of choice often point to student success in private schools as evidence of greater educational productivity in the private sector.[15] However, private schools can establish admissions criteria, such as minimum test scores, whereas public schools, in general, must accept all students. Private schools also tend to serve students whose parents are more affluent and well educated. Thus, it is not immediately clear that differences in performance between public school students and private school students are a direct result of the delivery of education or due to differences in their background.

Numerous studies have examined this issue, notably an HS&B-based study by Coleman et al. in 1981. This work was widely criticized because it did not

[15] On average, private schools have higher standardized test scores, graduation rates, and college matriculation rates than do public schools.

adequately control for selection into school sector. Goldhaber (1996) and Figlio and Stone (1997) both use the NELS:88 data to tackle the selection issue in the context of public and private schools.[16] They find little or no difference in the effectiveness of public and private schools. Simulations of what would happen if a student were to switch school sectors indicate that, holding all else equal, the large difference between public and private schools in mean standardized test scores is accounted for primarily by student background, school resources, and student selection, rather than differences in how effectively schools use the resources that they have.

Other data would permit a simulation of this type but would not account for teacher- and class-level differences. A very similar procedure has also been used in a series of studies examining the effectiveness of ability grouping (tracking).[17]

Conclusion: Further Improvements in the Data

In this paper we have argued that recent improvements in longitudinal data collection permit researchers to better tackle important unresolved educational policy issues, such as the effects of class size and teacher preparation on student achievement and the effectiveness of private schools. The availability of more detailed data with students linked to teachers reduces the likelihood of aggregation and omitted variables biases, and the design of NELS:88 with student-, class-, teacher-, and school-level information permits the estimation of a broader class of statistical models than has previously been possible. Although large-scale controlled experiments are clearly preferable, we believe that the advances made in nonexperimental data collection and methodology over the past decade represent a substantive improvement.

[16] This methodology uses a statistical procedure known as the Heckman two-step method (Heckman 1979), which requires variables that affect choice of school sector but do not affect student achievement. NELS:88 contains a richer set of variables that potentially fulfill this requirement. See Figlio and Stone (1997) for a detailed discussion of this point.

[17] Although a number of small-scale experiments have suggested that placing students in classes of heterogeneous ability benefits all students, this conclusion is controversial. Large-scale nonexperimental studies using NELS:88 suggest the effects are not so simple. See, for example, Brewer et al. (1995) and Argys et al. (1996).

We believe a number of improvements could be made in future NCES longitudinal data to further reduce the potential for omitted variables and aggregation biases and allow for the estimation of more sophisticated statistical models. For example, one problem with NELS:88 is that students are only surveyed every two years. This means that data about the students' academic and other experiences in the intervening year are lost.[18] As noted above, this creates the potential for bias in any model of student achievement growth. Additionally, it would be useful for research purposes to have more detailed class-level information, particularly on the socioeconomic status of the students in each class. This is crucial to understanding peer effects. The ideal would be to continue to collect data on teachers and classes and test scores that correspond to the students' exposure to particular teachers and classes.

As we have stressed in this paper, the link between students and teachers is a critical addition to NELS:88 and should be maintained. However, this link would be even more useful if additional data about teacher background were available. For instance, the few studies that have had measures of teacher (verbal) ability, for example, in the form of a teacher test score or selectivity of undergraduate institution, have found it is related to student achievement (Coleman et al.1966; Ehrenberg and Brewer 1995; Ferguson 1991). However, collecting a teacher test score could be controversial. It would be necessary to obtain information on standardized tests that teachers have taken in the past (e.g., SAT, ACT, and NTE) either through transcript collection or by administering new tests that measure teacher knowledge and ability. Either option could be costly and time-consuming. A third alternative that is likely to be useful, but less costly, is to collect evaluations of individual teachers by the school principal (since the principal is already being surveyed).

Information on the year that teachers obtained their licensure and the state from which they obtained their licenses would also be quite useful. This would be a relatively low-cost addition—perhaps as little as one item on the teacher survey—but it would allow researchers to better identify the effects of state and institutional policies. This appears particularly important given that policymakers in many states have recently overhauled (or are considering changing) their licensure and/or teacher preparation requirements.[19]

[18] Given budgetary constraints, one option would be to collect an abbreviated set of student and teacher variables in the intervening years.

[19] One possibility is to collect information in such a way that future data can be linked to the *Schools and Staffing Survey*, which contains these types of data.

In considering possible additions to the data, we recognize that there are constraints on the amount of information that can be collected. Thus, we would de-emphasize the importance of collecting items relating to student, parent, and teacher beliefs, attitudes, and feelings. We consider these to be intermediate variables that are, in many cases, byproducts of actual individual qualities and institutional policies. The rationale for this suggestion is that policymakers can only indirectly influence beliefs, attitudes, and feelings through the incentive structures they create (e.g., the structure of teacher compensation) through policy levers.

A major cost saving could also be achieved by putting less stress on collecting nationally representative samples. Sampling fewer schools with more data on students and classes within a smaller number of schools is an alternative.

We recognize that NCES has an obligation to provide national educational indicators; however, this is less important for the kinds of multivariate statistical models that researchers find most persuasive in tackling the most important policy questions. The reason is that, for statistical purposes, it is not necessary to have a nationally representative sample to obtain accurate estimates of the effects of the variables of interest.

Finally, one weakness of NELS:88 is the limited number of student observations per teacher and class. This limitation means that it is difficult to separate teacher effects from class-level effects. The more student observations per class and the more classes per teacher, the more we can learn about how teacher characteristics or behavior affects student outcomes and how these factors are related to the types of students being taught. While our findings on teacher preparation derived from the NELS:88 study are important for policy, future data collection could allow researchers to gain a more comprehensive understanding of the complex relationships between students, teachers, and schools.

References

Akerhielm, K. (1995). Does Class Size Matter? *Economics of Education Review 14*(3): 229–241.

Argys, L., Rees, D. I., and Brewer, D. J. (1996, fall). Detracking America's Schools: Equity at Zero Cost. *Journal of Policy Analysis and Management 15*(4).

Boardman, A. E., and Murnane, R. J. (1979). Using Panel Data to Improve Estimates of the Determinants of Educational Achievement. *Sociology of Education 52*: 113–121.

Brewer, D. J., Rees, D. I., and Argys, L. M. (1995, November). Detracking America's Schools: The Reform Without Cost? *Phi Delta Kappan 77*(3): 210–215.

Brewer, D. J., and Goldhaber, D. D. (1996). Educational Achievement and Teacher Qualifications: New Evidence from Microlevel Data. In B. Cooper and S. Speakman (Eds.), *Optimizing Education Resources* (pp. 389–410). Greenwich, CT: JAI Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* U.S. Department of Health, Education and Welfare. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1981). *Public and Private Schools*. *Report to NCES.* Chicago: National Opinion Research Center.

Ehrenberg, R. G., and Brewer, D. J. (1994). Do School and Teacher Characteristics Matter? Evidence from *High School and Beyond. Economics of Education Review 13*(1): 1–17.

Ehrenberg, R. G., and Brewer, D. J. (1995). Did Teachers' Verbal Ability and Race Matter in the 1960s? *Coleman* Revisited. *Economics of Education Review 14*(1): 1–23.

Ehrenberg, R. G., Goldhaber, D. D., and Brewer, D. J. (1995). Do Teachers' Race, Gender, and Ethnicity Matter? Evidence from NELS:88. *Industrial and Labor Relations Review 48*(3): 547–561.

Ferguson, R. (1991). Paying for Public Education: New Evidence on How and Why Money Matters. *Harvard Journal on Legislation 28*: 465–498.

Figlio, D., and Stone, J. (1997). School Choice and Student Performance: Are Private Schools Really Better? Unpublished paper. Department of Economics, University of Oregon.

Gamoran, A. (1996). Student Achievement in Public Magnet, Public Comprehensive, and Private City High Schools. *Educational Evaluation and Policy Analysis 18*(1): 1–18.

Goldhaber, D. D. (1996). Public and Private High Schools: Is School Choice an Answer to the Productivity Problem? *Economics of Education Review 15*(2): 93–109.

Goldhaber, D. D., and Brewer, D. J. (1997a). Evaluating the Effect of Teacher Degree Level on Educational Performance. In W. Fowler (Ed.), *Developments in School Finance, 1996* (NCES 97–535) (pp. 199–208). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Goldhaber, D. D., and Brewer, D. J. (1997b). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *Journal of Human Resources 32*(3): 505–523.

Goldhaber, D. D., Brewer, D. J., and Anderson, D. (1999). A Three-Way Error Components Analysis of Educational Productivity. *Education Economics 7*(3): 199–208.

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Education Production Functions. *Journal of Human Resources 14*(3): 351–388.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in the Public Schools. *Journal of Economic Literature XXIV*(3): 1141–1178.

Hanushek, E. A., Rivkin, S., and Taylor, L. (1996). Aggregation and the Estimated Effects of School Resources. *Review of Economics and Statistics 78*: 611–627.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1): 153–161.

Hedges, L.V., Laine, R., and Greenwald, R. (1994). A Meta-analysis of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher 23*(3): 5–14.

Lee, V. E., and Smith, J. (1997). High School Size: Which Works Best and for Whom? *Educational Evaluation and Policy Analysis 19*(3): 205–227.

Lee, V. E., Smith, J., and Croninger, R. (1997). How High School Organization Influences the Equitable Distribution of Learning in Mathematics and Science. *Sociology of Education 70*: 128–150.

Mason, W. M. (1995). Comment. *Journal of Educational and Behavioral Statistics 20*(2): 221–227.

Monk, D. H. (1992). Educational Productivity Research: An Update and Assessment of Its Role in Education Finance Reform. *Educational Evaluation and Policy Analysis 14*: 307–332.

Monk, D. H., and King, J. (1994). Multi-level Teacher Resource Effects on Pupil Performance in Secondary Mathematics and Science: The Role of Teacher Subject Matter Preparation. In R. Ehrenberg (Ed.), *Contemporary Policy Issues: Choices and Consequences in Education*. Ithaca, NY: ILR Press.

Parrish, T., and Brewer, D. J. (1998, August). *The Costs of Enrollment Increases and Class Size Reduction.* Report to U.S. Congress. Palo Alto, CA: American Institutes for Research.

Rock, D. A., and Pollock, J. M. (1991). *Psychometric Report for NELS:88 Base Year Test Battery.* Washington, DC: National Center for Education Statistics.

School-level Correlates of Reading and Mathematics Achievement in Public Schools

Donald McLaughlin and Gili Drori
American Institutes for Research

Introduction

The Schools and Staffing Survey (SASS) conducted by the National Center for Education Statistics (NCES) offers the most comprehensive picture available of the education system in the United States. Initiated in 1987–88 and repeated in 1990–91 and 1993–94, SASS consists of surveys of districts, schools, principals, and teachers that are associated with a national sample of schools. It offers information on issues such as policies, programs, services, staffing, and enrollment at both the district and the school levels, as well as the principals' and teachers' background, training, experience, perceptions, and attitudes. Given the broad reach of SASS, it can speak to a variety of important educational research and policy questions. The value of SASS would be even greater, however, if the relationship between these measures and the level of achievement in schools were known. As noted by others (Boruch and Terhanian 1996, Kaufman 1996), combining this survey information with data from other sources would allow SASS to more meaningfully inform debates over which factors relate to school effectiveness and could contribute to a broad-based evaluation of school improvement strategies.

The aim of this paper is to show the potential value of a linkage between the SASS database and information on student academic achievement collected by individual states.[1] Most states currently collect state assessment data[2] on their public schools and thus offer state-specific information on school performance in terms of student test scores. Although many different assessment instruments are used across the states, they all aim to provide an indication of the reading and mathematics achievement levels of their schools. By transforming each school's score to a *z score* relative to other schools tested at the same grade in a state, there is potential for pooling analytical results across states to increase both power and generalizability.

While pooling information from individual states into a single database can add substantial power to analyses identifying school-level correlates of achievement in SASS schools, it does not capture between-state sources of covariation with achievement. State policies and state demographics frequently limit the variation of education practices in a state, so that within-state associations with achievement are attenuated. To capture the full range of achievement variation between schools across the nation, one must include between-state variation. That variation is reported by State NAEP, the component of the National Assessment of Educational Progress (NAEP) which focuses on state-by-state achievement assessment. To evaluate whether combining state assessment scores with State NAEP data would increase the value

[1] This paper is based in part on a project carried out by the authors for the National Center for Education Statistics, through the Education Statistics Services Institute. The authors wish to thank the project officer for that project, J. Michael Ross, for his thoughtful suggestions and encouragement during the course of the work. We also wish to thank Adam Gamoran, Robert Hauser, Valerie Lee, and Stephen Raudenbush for their thoughtful reviews of an earlier draft. Nevertheless, the conclusions expressed in this paper are solely those of the authors, and no endorsement of these conclusions by reviewers or by the Center should be inferred.

We appreciate the help provided by the Center and 20 State Education Agencies in providing the data for this project, and we appreciate the work of staff of the American Institutes for Research, including Mary Anne Arcilla, Grace Wu, Elizabeth Hartka, and Inna Shapotina, in putting the database together. Finally, we recognize the thousands of hours of time spent by the respondents to these surveys and assessments, who provided information that can be used to improve education policies and practices.

[2] In 1994–95, 45 U.S. states had a statewide assessment system; the remaining five states either did not have a statewide system at all or had temporarily suspended their programs (National Education Goals Panel 1996). In 1995–96, 46 states administered statewide assessments (Roeber, Bond, and Braskamp 1997). In 1996–97, 45 states administered statewide assessments (Roeber, Bond, and Connealy 1998). Some educational assessment is done in every state, and in most of the few states without statewide testing programs, most districts use nationally standardized tests for assessment.

of the linked database, this study focused on states which both conducted statewide assessments in 1993–94 and participated in the 1994 State NAEP fourth grade reading assessment. Thus, in each of the 20 states included in this study, individual state assessment data are available for most or all public schools, along with summary NAEP results based on approximately 2,500 students in 100 random schools in the state.

To assess the potential value of combining SASS with achievement data at the school level requires three steps: (1) matching the schools on the 1993–94 SASS file with state reading and mathematics assessment scores for public schools in 20 states; (2) creating a comparable achievement measure for the matched schools from the combination of state assessment and State NAEP information; and (3) carrying out analyses to test hypotheses by modeling the relationship between a variety of SASS school-level responses and average student assessment scores at the school level. The hypotheses selected for the third step concern the identification of school-level correlates of student achievement. Although analyses of school-level information collected at one point in time cannot be used either to identify individual-level correlates of achievement or to draw direct causal inferences, they can potentially provide a basis of evidence for addressing issues of strategies for school improvement.

Combining SASS and State Assessment Data

The linkage of state assessment data to the SASS file required access to restricted information concerning the identification of SASS schools. NCES has established clear criteria for acceptable procedures for storing and using confidential information, and the American Institutes for Research (AIR) have complied with these criteria. Although the linkage might have been possible with information about schools' names and addresses, it was greatly facilitated by the use of an intermediate linkage of both SASS and state assessments through the Common Core of Data (CCD). The 1991–92 CCD file, which served as the sampling frame for SASS, identifies most of the 86,287 public schools in the country by both their federal and their state identification codes.[3]

[3] Information about CCD can be found on the NCES Web page: http://www.ed.gov/NCES/.

In most cases, school records on state assessment files included the state's identification code, which enabled an automated matching procedure.[4]

These 20 states contained 3,785 of the 8,767 SASS public schools. Of these 3,785 SASS public schools, 2,916 had students enrolled in grades corresponding to the state assessment; and 2,628 were identified as having both SASS and state assessment information.[5] Of these, 66 had no teacher data, and one had erroneous mean achievement scores, so the final file used for analysis contained 2,561 school records: 1,123 elementary schools, 496 middle schools, 595 secondary schools, and 347 combined-grade schools. The database includes at least 50 schools in each state and constitutes a broad sample of large and small, urban and rural, affluent and impoverished public schools.

The coverage of the range of educational contexts in the United States by the schools in the sample determines the extent to which inferences based on analyses of the database can be expected to generalize to other schools in the country. Although SASS includes both public and private schools, state assessment data are collected for public schools only; hence, the SASS student achievement subfile created for this report is limited to information on public schools.

The 2,561 public schools included in the study are only slightly different from the general population of American public schools on most measures examined. Although 51 percent of the elementary, middle, and secondary schools in the study sample were elementary schools, compared to 63 percent in

[4] A few of the schools were identified "manually" by matching their state, city, or zip code, either because the federal identification code was missing from the restricted SASS file or because the state identification code was not included on the assessment file. Details of the file development process can be found in Wu, Royal, and McLaughlin (1997).

[5] Of the other 288 SASS schools, 254 did not match with state assessment, and 34 merged with state assessment files but did not have usable mean scores in both reading and mathematics. In addition, 112 of the 254 nonmatching schools were special, alternative, or vocational education schools, or schools that had an enrollment of fewer then 10 in the grade assessed; and 86 were not included in one state's (Pennsylvania's) assessment sample in 1993–94. As a measure of the success of the matching process, 2,662 of the 2,718 SASS public schools that were expected to have matching assessment scores were matched, for a match rate of 97.9 percent.

the nation[6], the percentages of central city, urban fringe, and rural schools in the study sample were each within one percent of the percentages in the population.[7]

Results of analyses carried out separately for this sample of elementary, middle, and high schools, while not quantitatively representative of public schools in the nation, can suggest possible generalizations to other American schools. In any case, separate analyses by grade level are essential in using the SASS student achievement subfile, not only because different factors are related to achievement at different levels, but also because different achievement measures are used in each state at the different school levels

Finally, because the sample consists of a nonrandom subset of 20 states, no claims can be made that estimates of effect-sizes are quantitatively representative of the nation. The states included in the file are Alabama, California, Delaware, Florida, Georgia, Hawaii, Kentucky, Louisiana, Maine, Massachusetts, Michigan, Montana, New Hampshire, New York, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, and Washington.

A School-level Measure of Student Achievement

The first step in rendering state assessment scores comparable is to compute each school's score as it relates to other schools' scores in the state at the same grade. That is, the (unweighted) mean score of the schools in the database for the particular grade and state is subtracted from each of the scores to create a score with a mean of 0 at each state and grade; then these scores are divided by the standard deviation of the school scores to create a score with a standard deviation of 1.0.

Using this measure, third grade reading scores in one state, fourth grade reading scores in another, and fifth grade reading scores in a third are taken to be comparable achievement measures for the purpose of computing within-state correlations across elementary schools with factors such as average class

[6] The SASS student achievement subfile included 347 ungraded schools, or 14 percent of the total, a much larger representation of ungraded schools than in the population (3 percent). However, ungraded schools were not included in the analyses reported in this study.

[7] Additional descriptive comparisons of the schools in the SASS student achievement subfile with other public schools are reported by Wu, Royal, and McLaughlin (1997).

size and school behavioral climate. All achievement differences between states are removed in this measure, so comparisons with school characteristics would also need to remove between-state variation in school characteristics.

The second step is, therefore, to re-introduce between-state variation using a common standard, State NAEP. In a separate study, the State NAEP schools were linked to state assessment scores; and the means, standard deviations, and correlations of State NAEP school means with state assessment school means were computed. Those results were used in this study to create reading and mathematics achievement measures that include (1) between-state variation in means, (2) between-school variation proportional to between-school variation among State NAEP schools, and (3) a factor that attenuates within-state variation for states in which the assessment is only moderately correlated with State NAEP. The effect of the third factor, multiplying by the correlation between the NAEP and state reading assessment scores, "projects" the state assessment variation onto the NAEP scale, capturing that part of the state assessment score that is like NAEP.[8]

Thus, this achievement score "spreads schools apart" in states in which (a) school NAEP scores are more varied and (b) the state assessment appears to be measuring skills highly related to NAEP. The effect of this spreading is to give greater weight to variations within these states in the estimation of correlations of achievement with SASS measures.

Important assumptions are needed to apply the NAEP adjustment to the scores at grades other than fourth and to mathematics scores. The between-state NAEP adjustment was based on the 1994 State NAEP grade 4 assessment for reading and on the 1992 and 1996 State NAEP grades 4 and 8 assessments for mathematics. Application of these adjustments to state assessment scores in middle school (for reading) and high school is based on the assumption that variation in achievement *between states* is stable across grades. The 1992 and 1996 State NAEP mathematics assessment results support this assumption, in that the correlations between the grade 4 and grade 8 state means are 0.95 and 0.92, respectively (Mullis, Campbell, and Farstrup 1993; Reese et al. 1997).

[8] An alternative to projecting the scores onto the NAEP reading scale would be to omit the third factor, treating each state's reading assessment as the relevant outcome for that state. However, it would be more difficult to characterize an achievement measure that is based on different scales across states.

However, no information is available regarding State NAEP means at the high school level.

The use of 1992 and 1996 State NAEP state means and standard deviations to construct the 1994 mathematics adjustment assumes that state means varied smoothly, if at all, from 1992 to 1994 to 1996. In fact, the correlations of this study's 20 State NAEP mathematics means between 1992 and 1996 were 0.91 for grade 4 and 0.93 for grade 8, suggesting that interpolating 1994 figures from 1992 and 1996 figures is reasonable.

Finally, the correlations used in the adjustments were based on grade 4 reading assessments. Use of these correlations in the adjustment of within-state variation in math scores assumes that between-state variations in reading and math are highly correlated. Because state assessments usually combine reading and math tests from the same publisher and in the same testing session, it is plausible to assume that factors that would affect the reading correlations in different states (e.g., the reliability of the state assessment instrument and distribution of state assessment scores) would also affect the math correlations. The 1992 State NAEP assessments in reading and mathematics in grade 4 indirectly support this assumption, in that the correlation between reading and mathematics state means is 0.94 (Mullis et al. 1993). Nevertheless, the question remains whether the results of substantive analyses will be diminished by the extrapolation of between-state variation in average achievement from grade 4 to middle and high school variation. Comparative analyses of NAEP-adjusted vs. pooled within-state findings across school levels (carried out in this study) address this question.

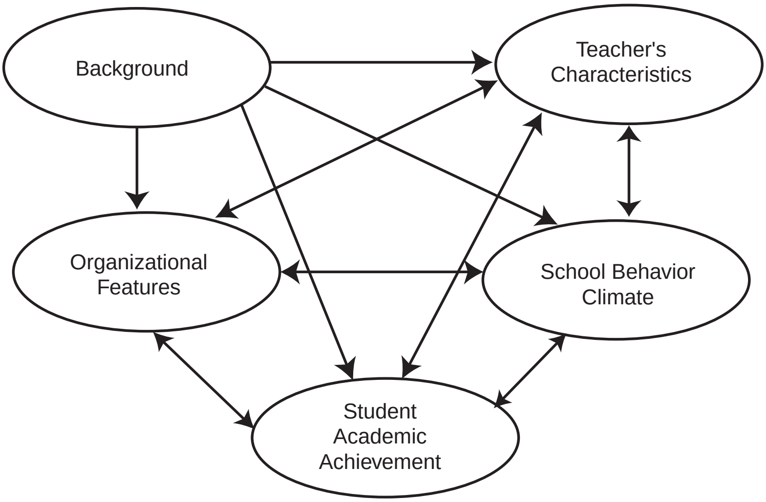
Although State NAEP data were used to capture between-state variation in achievement, it would be highly misleading to interpret the SASS student achievement measures as a surrogate of the school's average NAEP proficiency. First, State NAEP differs from individual state assessments in student sampling (each student takes only a fraction of the NAEP test), administration (a federal government contractor trains test administrators and monitors many testing sessions), motivation (NAEP is a low-stakes assessment with no individual student or school reporting), and item formats (NAEP has a substantial portion of extended open-ended items). The achievement measure developed for this study yields an unbiased estimate of school NAEP means, but with different standard errors in each state. The measure is not dependent on evidence that NAEP is *equated* to the various state assessments, and evidence that they might be “equatable” (i.e., parallel) was not sought. In fact, it is highly unlikely that the states’ individual assessments are parallel to NAEP, due to differences in administration, item format, and content frameworks. In other research, we have shown that it is feasible to project state assessment results onto the NAEP scale without assuming that the tests are parallel (McLaughlin 1998). Second, the correlations between NAEP and state assessments differ substantially between states. Although the median correlation in these states in 1994 was 0.70, the smallest three correlations were between 0.30 and 0.50 (Wu, Royal, and McLaughlin 1997). Within-state variation of these synthetic NAEP school means will be smaller than variation of actual NAEP school means, especially in states where assessments are not highly correlated with NAEP.

School-level Correlates of Achievement

Student academic performance is shaped by multiple factors relating to the school, teaching process, students’ social and family backgrounds, and community; also, a school’s reputation for academic performance can affect parents’ decisions, students’ behavior, and teachers’ attitudes and decisions. We model student achievement in American public schools as related to four types of factors: (a) students’ background, (b) four organizational features of the school, (c) professional characteristics of the teachers, and (d) school behavior climate. While all these factors affect student academic success, they also interact with each other; and organizational characteristics and teacher choices can be affected by achievement at the school. We therefore conceptualize the interrelationships among the five categories in this model as a web of interactions. Figure 1 graphically describes the general model.

The model shown in figure 1 refers to the school as a unit. Of course, achievement is an individual student characteristic, and the majority of variation in achievement is among students in the same classroom and between classrooms in the same school. Nevertheless, there is substantial reliable variation in achievement between schools; and from a policy perspective, there may be actions that can improve the overall achievement level in a school. While analyses at the school level may not shed light on individual variation in learning, they can provide evidence on the correlation between school reform policies and achievement. Hierarchically structured data, with both individual student data and schools and staffing data, such as collected in NAEP and NELS:88, facilitate understanding of the correlates of individual student achievement;

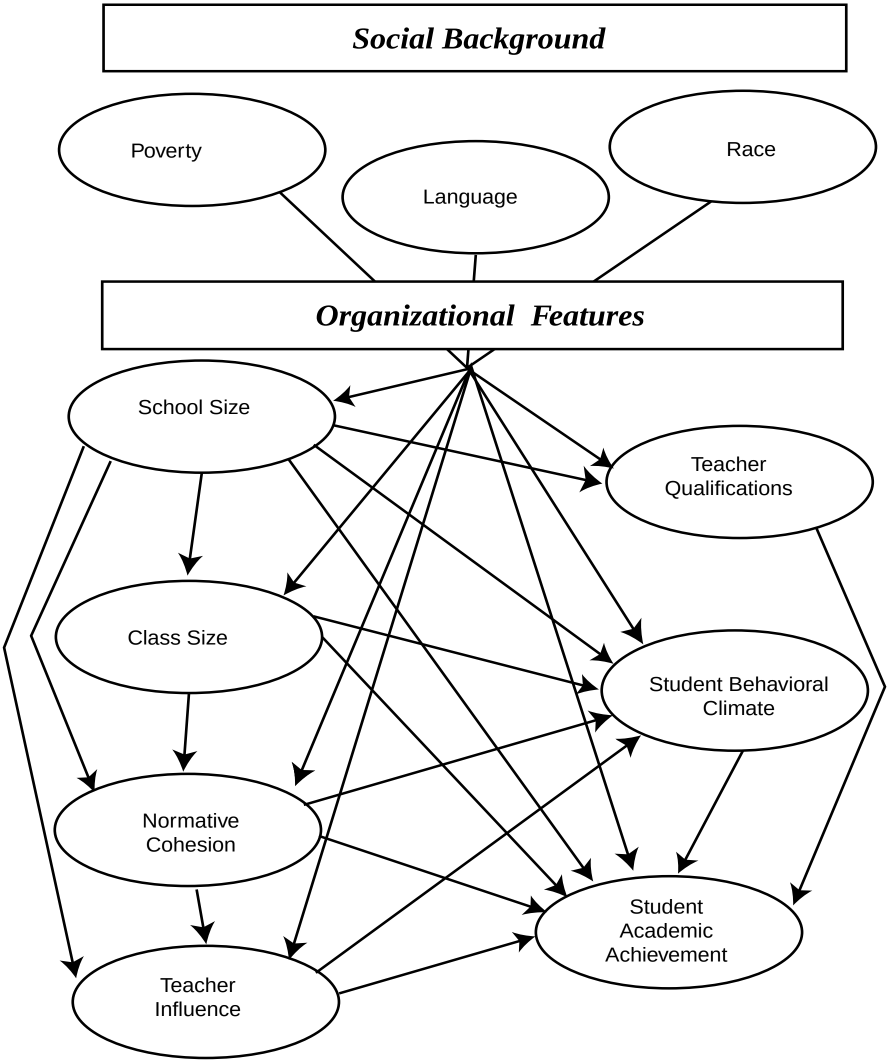
Figure 1. School-level Correlates of Student Achievement: General Model



but these data are much more costly to collect on a school sample the size of SASS than is the construction of a synthetic achievement measure from existing NAEP and state assessment data. In any case, the existence of within-school variability does not threaten the validity of analyses at the school level.

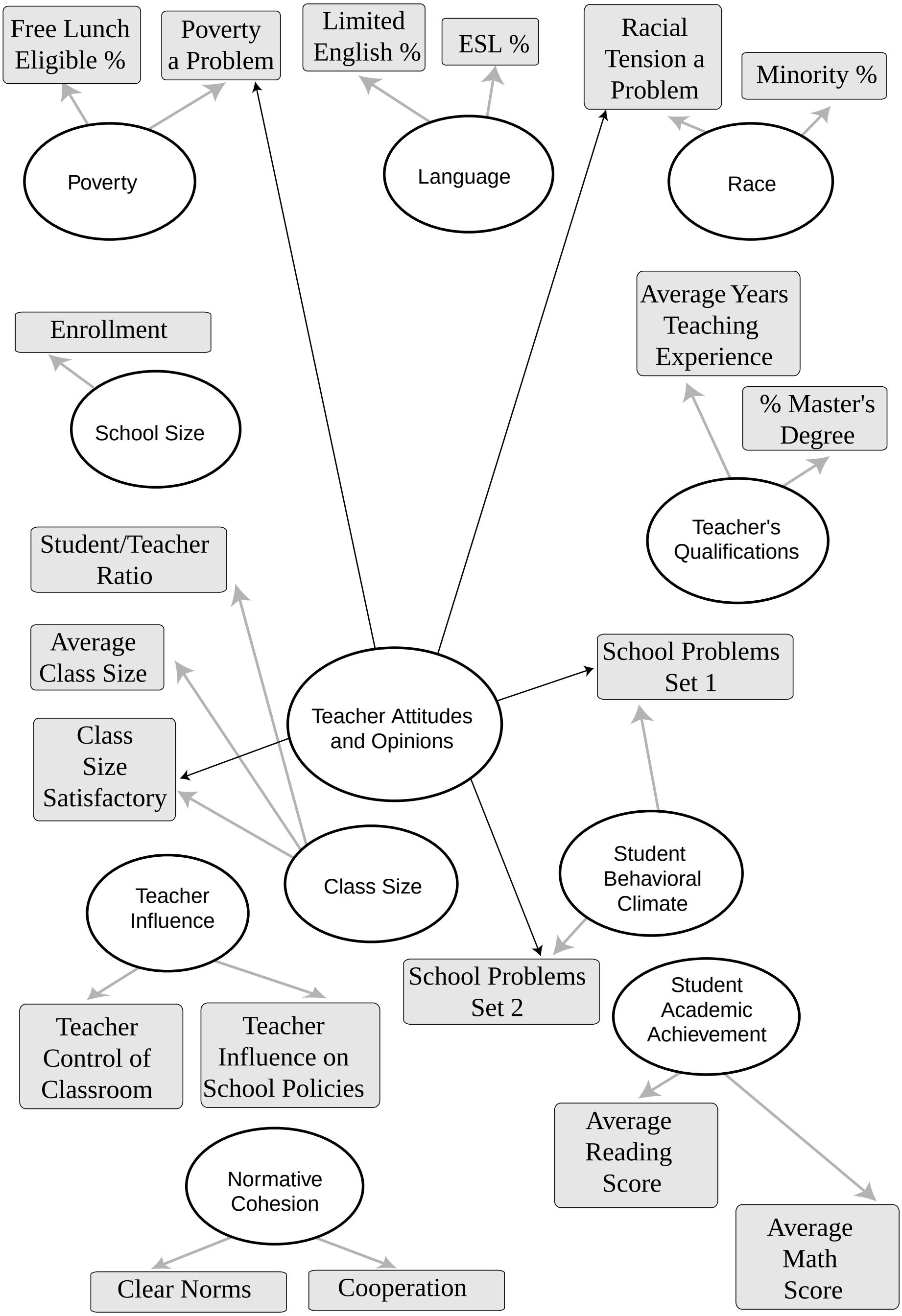
Using the SASS student achievement subfile, the model in figure 1 can be further specified in a variety of ways, one of which is shown in figure 2. The background category in the model is represented by three factors: (1) percentages of students in poverty, (2) percentages with language barriers, and (3) percentages in racial-ethnic minorities. The organizational category is represented by four factors: (1) school size (total enrollment), (2) average class size, (3) teachers' perceived influence, and (4) normative cohesion. The aim of the analyses to be carried out is to test hypotheses about the relations among these factors, either in terms of partial correlations or in terms of fits of linear models, such as that represented graphically in figure 2. By testing these models, inferences can be made about correlates of achievement across a wide range of public schools, although the "arrows" cannot, in most cases, be taken to indicate a direction of causality because of the alternative explanations of many of the correlations.

Figure 2. One Possible Path Model Relating Achievement and School and Background Factors



Each of the factors in the model, except school size, is represented by multiple measures in SASS, as indicated in figure 3. In structural equation terminology, figure 3 presents the measurement model corresponding to the structural model in figure 2. The arrows in figure 3 indicate assumptions about the sources of variance in the observed measures. Each of the factors in figure 2, except school size, is represented by at least two indicators, providing the

Figure 3. Measurement Model for School-level Correlates of Achievement



capacity for estimating the contribution of measurement error to variance in the indicators. Also indicated in figure 3 is a factor, *teachers' attitudes and opinions*, representing a common response pattern among five of the measures. These measures may be more positively intercorrelated than other measures because they all represent teachers' subjective opinions about their schools.

The two indicators of school behavior climate are based on 20 items concerning teachers' perceptions of problems in the school. The two parallel measures were constructed by averaging balanced halves of these items. For example, drug abuse is in one set, alcohol abuse in the other; student absenteeism in one and dropping out in the other; vandalism in one and robbery in the other. Two topics for which there were multiple items, tardiness and attacks on teachers, were included in both sets.

Determining the correlates separately for elementary, middle, and high schools provides an opportunity to explore the patterns of change in the correlations over the school years. Much like Herriot and Firestone's (1984) arguments that the images of schools vary by level[9] and that each school level operates differently,[10] we anticipate that the relative importance of the various factors to student academic achievement will vary by school level. For example, we anticipate that the relationship between social background factors and student performance will be found to be greater in elementary schools than in secondary schools.

Student Background

Family background and socioeconomic status are consistently shown to be related to student achievement (Coleman et al. 1966; Hanushek 1986). While the major purpose of research on schools is to identify characteristics of *schools* that contribute to student achievement, it is essential to take background characteristics into account because they affect the intercorrelations of school measures. For example, suppose that poor children were found to be attending schools with chipping paint. We would expect to find a correla-

[9] Elementary schools are imagined to be more rational and bureaucratic, while high schools are seen as more anarchic and envisioned to be a loosely coupled system.

[10] In elementary schools, curriculum is more limited, while in high schools, curriculum is broad. Also, the operations of elementary schools are more centralized and consensus-driven, while those in high schools have high levels of complexity and differentiation.

tion between chipping paint and test scores, but one should not infer from that correlation that painting schools will improve test scores. It is the pattern of correlates among schools with students with similar backgrounds that is of interest. One need not address social issues about the sources of the impact of background variations on achievement in school to realize the need to control for these factors in modeling school processes.

It is particularly important to include background factors in this study of school-level factors, because the database does not contain longitudinal achievement data on individual students or student cohorts. School characteristics and policies are more likely to be correlated with school-level variation in gains in achievement than with achievement differences measured at one point in a student's career. Including student background measures in the model serves to control for much of the between-school variation in achievement potential that students bring to school. Nevertheless, the lack of "pre" measures of achievement underlines the need to interpret the results of analyses of SASS data relevant to the model in figures 1 and 2 as causally indeterminate.

Differences in the family backgrounds of students in different schools are reflected in three factors: (1) poverty, (2) English language proficiency, and (3) racial-ethnic minority status. The level of *poverty* in a school is measured by two indicators available from the SASS database: (1) the percentage of students who qualify for the national school lunch program and (2) teachers' identification of poverty as a problem. The level of *English language proficiency* in the school is composed of two similar indicators: (1) the percentage of students identified as Limited English Proficient (LEP) and (2) the percentage of students participating in the school's English-as-a-Second-Language (ESL) program. Finally, the *minority status* of a school is measured by (1) the percentage of white students in the school and (2) teachers' identification of racial tension as a problem.

A unique strength of the SASS database is that background factors can be measured by an objective indicator, the percentage of students with the corresponding characteristic, and by the perception of a sample of teachers in the school that said factor is a source of problems in the school. For example, the underlying factor of poverty is only imperfectly measured by the percentage of children who are eligible for the federal free lunch program, due to differences in cost of living and in eligibility counting procedures between school districts. Incorporating the perceptions of a sample of teachers in the school that poverty is a problem in the school can eliminate some of the error of measurement of poverty. Of course, there are different kinds of error in teacher perceptions, ranging from different interpretations of survey items to sampling error, but the combination of the two indicators can be expected to have greater validity for the impact of poverty on learning in the school than either separately. Because SASS has many related objective and subjective measurements, one can control for measurement error associated with variation in teachers' use of the response scale (e.g., some teachers mean something more serious by "problem" than others do) by estimating the extent to which each teacher tends to be a positive or negative responder to attitude and opinion items.

Alternative indicators of a particular factor, such as subjective and objective assessments of poverty, must be correlated, but they need not be highly intercorrelated; however, a low intercorrelation is likely to limit the power of the data to measure correlations with achievement. The intercorrelations of the indicators included in each composite factor are shown in appendix A following this paper. For example, for poverty, the intercorrelations between the objective and subjective indicators are 0.55, 0.51, and 0.47 for elementary, middle, and secondary schools, respectively.[11] These intercorrelations are themselves limited by the "reliability" of the indicators that are based on averages of teachers' responses. If different teachers see the same school very differently, the average of their responses cannot tell a great deal about the school, per se. The percentage of variance in school means that is attributable to between-school variation, which is a measure of this reliability, is also shown in appendix A. For the average rating of whether poverty is a moderate or serious problem, the reliabilities are 0.73, 0.79, and 0.84 for teachers in elementary, middle, and high schools, respectively.

School Organizational Features

Both objective features of a school's organization such as its class sizes and subjective features such as the level of cooperation among its staff may be correlated with achievement. However, unlike the social background factors,

[11] The corresponding intercorrelations for the minority and language factors are approximately 0.45 and 0.75. The language factor, unlike the other background factors used here, is defined by two objective indicators.

these features are endogenous and, to a varying extent, under control of the principals and teachers in the school. For example, a magnet school with a reputation for attracting students with potential for careers in science is likely to have higher test scores than other schools, purely as a function of the backgrounds of the students who enroll; and its magnet status may affect class sizes, either enlarging them to respond to demand or lowering them as a result of special funding as a magnet school. Although the specific models presented in this report focus on accounting for variation in achievement in terms of variation in organizational characteristics, the direction of causality is not determined in these data.

SASS has a wide range of information about schools obtained from the principals in the administrator and school questionnaire and from the teachers in the teacher questionnaire. Four organizational features have been selected for inclusion in the model for this report: (1) school size, (2) average class size, (3) teachers' sense of their influence over school affairs, and (4) normative cohesion of the school's staff. A variety of other SASS organizational measures, such as organizational complexity as reflected in the number of different kinds of positions in the school, organizational goals as expressed by the principal, perceptions of outside influence on decisions by state agencies and local school boards, and staff diversity, might be included in a more elaborate model.

School size, as measured by total enrollment, has been shown to have a significant effect on the school's performance, yet the direction of the effect has not been consistent. Because school size has different implications for instructional resources at the elementary and secondary levels, and because school size is highly correlated with the sizes of classes in the school, the correlations of school size and achievement measures vary with the type of schools studied and the variables included in the study.

Class size, according to common sense, should have an effect on students' academic performance. However, it has proven difficult to isolate and demonstrate that effect. SASS has three indicators of a school's average class size: (1) average class size as reported by the sampled teachers;[12] (2) the school's

[12] For a teacher teaching a single self-contained classroom, class size is the total number of students enrolled in the teacher's class at the school. By contrast, for a teacher teaching multiple departmentalized classrooms, class size is calculated as the average number of students in the teacher's classes.

teacher/student ratio, calculated as the total number of students in school divided by the sum of the full-time teachers in the school and one-half the number of part-time teachers in the school; and (3) the average of the sampled teachers' ratings of satisfaction with the size of their classes.

As a measure of the extent to which the classrooms in a school have too many students for optimal learning, each of these has a different source of measurement error, such as variation in staff counting methods, teacher sampling variation, invisibility of class size remedies (e.g., teacher aides and parent volunteers), and different criteria for teachers' satisfaction. Although none of these by itself is a perfect measure of a school's class sizes, their combination may provide a more valid measure of the impact of class size on student learning than any of them considered separately.[13]

Teachers' sense of influence, measured by the average of sampled teachers' responses to SASS questions about their perceptions that they have influence over school policies[14] and over matters concerning their own class[15], differentiates schools with varying management styles and teacher roles. Although this factor may not have a measurable direct effect on achievement, the sense of efficacy represented by this factor may be related to the general climate in the classroom, which in turn can affect learning.

Finally, *normative cohesion* of the staff refers to the cultural solidarity among staff members in the school or the collective norms that govern staff behavior in this organization and may also be correlated with the climate in classrooms in the school. Normative cohesion can be measured by two SASS

[13] Among elementary schools, the reliability estimate of teacher-reported class sizes as an indicator of the school is low, 0.31, and the intercorrelations of that with the other class size indicators are also low, 0.28 and 0.29. This may limit the potential elementary school correlations of class size with achievement.

[14] Included items refer to influence over school discipline policies, the content of inservice programs, hiring new full-time teachers, deciding priorities in spending the school budget, evaluating teachers, and establishing a curriculum. The reliability of the school mean is about 0.62.

[15] Included items refer to control in one's own classroom over selection of textbooks and other instructional materials, selection of contents, topics, and skills to be taught, selection of teaching techniques, evaluating and grading students, disciplining students, and determining the amount of homework to be assigned. The reliability of the school mean ranges from 0.36 at the elementary level to 0.50 at the secondary level.

composite measures: (1) a score for the clarity of norms[16] and (2) a score for staff cooperation.[17]

Teachers' Qualifications

Common sense leads to the expectation that more qualified teachers create more effective learning environments in their classrooms, so part of the variation in achievement between schools may be due to teachers' qualifications. In SASS, a variety of teachers' characteristics are recorded for samples of five teachers per school, on average. Because level of education and amount of teaching experience are widely used to determine the pay scales of teachers, these form a logical basis for measuring teacher qualifications. In particular, the school level measures are (1) average years of teaching experience and (2) the percentage of teachers who acquired at least a master's degree.

Although teachers' qualifications are included in the analysis, the factor is relatively weak, compared to the other factors.[18] A wide variety of other SASS measures of teacher qualifications should be included in a study focusing particularly on teaching quality and achievement, including out-of-field teaching, selectivity of the colleges the teachers attended, amount of inservice training, hours spent on school-related activities, and training to teach Limited English Proficient (LEP) students.

[16] Calculated as the average of teachers' responses to SASS questions about sharing beliefs and values with colleagues in school, receiving support from parents, goals and priorities in school being clear, rules being consistently enforced by all teachers, principal backing up staff members, principal letting staff know what is expected of them, and principal having a clear vision of the type of school wanted and communicating this model to staff. The mean correlation among these seven measures is 0.42, and the reliability of the school mean is about 0.57.

[17] Calculated as the average teacher's responses to SASS questions about getting cooperative effort among staff members, making a conscious effort to coordinate course content with other teachers, planning with media specialist or librarian an integration of their specialty into teaching, and viewing the behavior of school administration as being supportive and encouraging. The mean correlation among these four measures is 0.20, and the reliability of the school mean is about 0.48.

[18] There is substantial within-school variation in years of teaching experience (reliabilities of 0.29, 0.30, and 0.37 at the three grade levels); correlations with the percent of responding teachers with a master's degree range from 0.14 to 0.27.

The relation between a school's averages of teachers' qualifications and achievement is complex. It may reflect choices by teachers of where to teach and of school districts as to how to allocate resources, as well as the impact of experience and training on effective learning environments. As with school organizational factors, the direction of causality of the relations between teacher qualifications and achievement cannot be inferred directly from correlations, but these correlations provide valuable evidence of relations that must be explained in some manner.

School Climate

Student achievement is difficult when the climate in a school reflects factors such as drugs, violence, vandalism, truancy, lack of respect for teachers, and lack of enthusiasm for learning. These characteristics are difficult to measure in a uniform manner across a large sample of schools, but SASS has attempted such a measurement by asking teachers to indicate which of two dozen different types of potential problems are serious, moderate, minor, or nonexistent in their schools. To assess the extent of measurement error in these perceptions, two composite measures can be constructed by arbitrarily dividing the problem ratings into halves.[19]

Although a positive relationship between school climate and achievement can be expected, the direction of causality in this relation is particularly ambiguous. There is likely to be a positive feedback between students' focus on achievement and teachers' perceptions of their behaviors and demeanors. Nevertheless, evidence about the significance of this correlation, and of its mediating role in other relations with achievement in schools, is valuable.

Analytical Method

The SASS student achievement database contains school-level statistics on hundreds of measures for over 2,000 public schools in 20 states. The potential for analyses of this database is enormous. In this report, we have constructed a set of 18 composites of SASS items; and for a baseline demonstration of analytical feasibility, the database has essentially been reduced to the intercorrelation matrix of these 18 composites, along with two assessment

[19] The reliabilities of the two climate composites range from 0.70 to 0.86 for the three school levels, and their intercorrelation ranges from 0.50 to 0.72.

scores, for schools at each of three grade levels, plus, for the purpose of standard error estimation, an indicator of the state in which each school resides.[20] However, the raw correlations between the 18 composites and the two achievement measures do not provide meaningfully interpretable information, because many of the apparent correlations are mere reflections of correlations among other measures, and other "real" correlations are masked by confounding measures and can only be uncovered by controlling for those confounders. Thus, the first meaningful stage of analysis is to examine partial correlations between achievement measures and SASS composite measures of school organization and climate and teacher qualifications, controlling for social background factors.

Partial correlation analysis, as a method of testing for significant relations between school composites and achievement measures, has the advantage that it is neutral with respect to causal ordering; but the picture of the correlates of achievement provided by partial correlations is limited, in that multivariate structure is not apparent. For example, the partial correlations of both normative cohesion and school climate with achievement may be positive, but one cannot discern from them whether this is due to a common factor in climate and cohesion or to independent factors. A form of multivariate regression is needed to identify the more complex structure.

The simplest form of multivariate regression is ordinary least squares (OLS) linear regression. This methodology can reveal the multivariate structure when its assumptions are satisfied, but an important assumption in OLS modeling that limits its value for educational research is that the "predictor" measures are measured without error. With many databases, this assumption is untestable, because only a single measure of each construct is available. However, the SASS database with its multiple sources of information about school-related constructs offers the opportunity to take measurement error into account when modeling the structural relations among factors. Structural equation modeling (SEM) jointly models the structural relations among latent factors (as in figure 2), while simultaneously modeling measurement error (as in fig-

[20] Although exploration of other functional forms may yield additional insights, estimation purely in terms of linear models is an efficient initial step, because software packages are readily available and most important relations among educational factors are monotonic and therefore "visible" to linear analyses.

ure 3). Computer programs such as LISREL, EQS, and SAS PROC CALIS can be used to estimate the variance components in SEM (see Bollen and Bollen 1989). The primary analytical results presented in this report are SEM analyses based on SAS PROC CALIS estimation.

SEM is particularly helpful in specifying, estimating, and testing hypothesized relationships among meaningful concepts, or factors,[21] by allowing such concepts to be estimated from several indicators, or measures. In SEM, the variance of a latent variable reflects the variation on the common factor among indicators of that latent variable, as measured by their intercorrelations. In SASS, the indicators themselves can be determined as composites of responses to individual items. For this study the construction of indicators was based on both judgment and factor analyses of items.

The structural model specifications for estimation of student achievement correlates corresponding to figure 2 are given by the following set of equations. A complementary measurement model relates each of the latent predictors to multiple SASS measures with the exceptions that reading achievement, mathematics achievement, and school size are each based on a single measure.

$$\begin{aligned}\eta_{\text{Reading}} = {} & \Gamma_{11}\,\eta_{\text{School climate}} + \Gamma_{12}\,\eta_{\text{Tchr qualifications}} + \Gamma_{13}\,\eta_{\text{Teacher control}} + \Gamma_{14}\,\eta_{\text{School size}} \\ & + \Gamma_{15}\,\eta_{\text{Class size}} + \Gamma_{16}\,\eta_{\text{Norm cohesion}} + \Gamma_{17}\,\xi_{\text{Language}} + \Gamma_{18}\,\xi_{\text{Poverty}} + \Gamma_{19}\,\xi_{\text{Race}} + \delta_1\end{aligned}$$

$$\begin{aligned}\eta_{\text{Mathematics}} = {} & \Gamma_{21}\,\eta_{\text{School climate}} + \Gamma_{22}\,\eta_{\text{Tchr qualifications}} + \Gamma_{23}\,\eta_{\text{Teacher control}} + \Gamma_{24}\,\eta_{\text{School size}} \\ & + \Gamma_{25}\,\eta_{\text{Class size}} + \Gamma_{26}\,\eta_{\text{Norm cohesion}} + \Gamma_{27}\,\xi_{\text{Language}} + \Gamma_{28}\,\xi_{\text{Poverty}} + \Gamma_{29}\,\xi_{\text{Race}} + \delta_2\end{aligned}$$

$$\begin{aligned}\eta_{\text{School climate}} = {} & \Gamma_{31}\,\eta_{\text{Tchr control}} + \Gamma_{32}\,\eta_{\text{School size}} + \Gamma_{33}\,\eta_{\text{Class size}} + \Gamma_{34}\,\eta_{\text{Norm cohesion}} \\ & + \Gamma_{35}\,\xi_{\text{Language}} + \Gamma_{36}\,\xi_{\text{Poverty}} + \Gamma_{37}\,\xi_{\text{Race}} + \delta_3\end{aligned}$$

$$\eta_{\text{Tchr control}} = \Gamma_{41}\eta_{\text{School size}} + \Gamma_{42}\,\eta_{\text{Norm cohesion}} + \Gamma_{43}\,\xi_{\text{Language}} + \Gamma_{44}\,\xi_{\text{Poverty}} + \Gamma_{45}\,\xi_{\text{Race}} + \delta_4$$

$$\eta_{\text{Tchr qualifications}} = \Gamma_{51}\,\eta_{\text{School size}} + \Gamma_{52}\,\xi_{\text{Language}} + \Gamma_{53}\,\xi_{\text{Poverty}} + \Gamma_{54}\,\xi_{\text{Race}} + \delta_5$$

$$\eta_{\text{Norm cohesion}} = \Gamma_{61}\eta_{\text{School size}} + \Gamma_{62}\,\xi_{\text{Language}} + \Gamma_{63}\,\xi_{\text{Poverty}} + \Gamma_{64}\,\xi_{\text{Race}} + \delta_6$$

$$\eta_{\text{Class size}} = \Gamma_{71}\,\eta_{\text{School size}} + \Gamma_{72}\,\xi_{\text{Language}} + \Gamma_{73}\,\xi_{\text{Poverty}} + \Gamma_{74}\,\xi_{\text{Race}} + \delta_7$$

$$\eta_{\text{School size}} = \Gamma_{81}\,\xi_{\text{Language}} + \Gamma_{82}\,\xi_{\text{Poverty}} + \Gamma_{83}\,\xi_{\text{Race}} + \delta_8$$

[21] Technically referred to as latent variables, yet also known as unobserved or unmeasured variables.

Because the data are correlational, it should be pointed out that the analysis is also consistent with views that characteristics like school climate, cohesive norms, cooperation, and satisfaction are affected by the kinds of skills and attitudes that children bring to the school, which are best reflected in achievement scores. For example, if the student peer norm is to focus on class work, there is likely to be less of a problem with absenteeism, tardiness, and class cuts. Likewise, a negative correlation between average class size and average achievement may indicate that smaller classes facilitate achievement, but it may also be due to a socioeconomic variation between school communities that affects both class size and achievement. Nevertheless, the methodology for the analyses is a variant of linear modeling with asymmetric "independent" and "dependent" variables. Therefore, although the correlational results may appear to be couched in terms of "effects" of some factors on others, these "effects" merely indicate the partitioning of the variance in the "dependent" factors into covariances with "independent" factors. This use of the term "effect" should not be confused with its use in the description of causal relations. Without carefully controlled experimental studies, the direction of causality cannot be inferred, only conjectured.

In addition to SEM analyses and in order to verify the robustness of the findings, we also carried out ordinary least squares (OLS) linear regression analyses.[22] For these analyses, we created composites for poverty, English language proficiency, racial-ethnic minority, teacher qualifications, school climate, school size, class size, teacher control, and normative cohesion, using the same measures as in the SEM analyses. Separate regressions were performed for elementary, middle, and secondary schools.

[22] The major difference between the two methodologies is in how they treat measurement error in predictive measures. The basic assumption of OLS linear regression is that predictors are measured "without error." That is, each estimated coefficient represents the extent to which a measure *per se*, not an underlying construct that it measures, accounts for variation in the dependent variable. SEM analysis, on the other hand, accounts for measurement error and allows for the specification of correlations among both the constructs and the measurement errors. On the other hand, SEM's greater flexibility must be weighed not only against the additional computational complexity, but also against the additional complexity of interpretation. Relations among latent variables do not have the same simplicity as relations among observable variables. Therefore, our approach is to examine and compare the results of both methodologies to identify robust patterns in the correlates of school-level achievement.

Finally, because this report was not focusing on differences in achievement correlates between states, the powerful analytical technique of hierarchical linear modeling (HLM) was not used. Hierarchical modeling would be a powerful tool in the use of these data to study the effects of state educational policies on school factors and achievement.

The value of the SASS student achievement subfile depends on whether meaningful patterns of statistically significant results emerge from analyses of relations between SASS measures and achievement. Thus, assessment of statistical significance is central to the study. Unfortunately, the usual tests of statistical significance available in common statistical packages are based on the assumptions of simple random sampling, and the SASS student achievement subfile is far from a simple random sample. Because of the similarity of the unweighted school sample to the universe of American public schools, unweighted analyses are appropriate.[23] However, the variance components within and between states cannot be expected to be uniform. Therefore, to provide valid estimates of the standard errors of statistical estimates, for the purpose of statistical significance testing, another method is needed. For this study, standard errors for all statistics were estimated by repeating each analysis on 100 random half-samples of schools. The standard deviation of the statistics computed for the various half-samples provided valid estimates of the corresponding standard errors. Because there is systematic variance in achievement measures between states (i.e., there is significant variation among NAEP mean state scores), it was necessary to select the random half-samples by state. That is, each half-sample consists of all the schools in the database for a randomly selected half of the 20 states. Although this method of standard error estimation is not as computationally efficient as balanced repeated replications, it provides valid standard error estimates.

Results

The major question addressed by these analyses of the SASS student achievement subfile is: Are organizational factors, teachers' qualifications, and

[23] To support reporting state-by-state statistics, SASS purposely (proportionately) oversamples schools in less populous states. As a result, differential weights are needed to estimate descriptive statistics for groups of states. Because the purpose of this study is not to produce descriptive statistics and because differential weighting substantially reduces the precision of estimates, differential weights are not used in this study.

school behavioral climate correlates of school mean assessment scores? The value of the SASS student achievement subfile is tested by the answers it gives to this question. Although the analyses reported here merely scratch the surface of the potential for analyses of these data, they should provide evidence of a meaningful pattern of relations between school-level factors and assessment scores.

The starting point for this research is the assumption that there is meaningful variation in assessment scores between schools. The choice of educational policies depends on the extent to which that variation is attributable to factors that are under a local school system's control, as compared to factors associated with the communities in which the schools are located. Estimation of the relative correlations of background and school-based factors with achievement is not straightforward, because background factors at least partially determine the levels of school-level factors. If, for example, teacher qualifications are correlated with student achievement, this may be due both to the fact that more qualified teachers teach more effectively and to the fact that higher achieving schools can attract more qualified teachers. Nevertheless, any analyses of school-based factors must control for background differences.

Three categories of school-based factors are included in these analyses: (1) school behavior climate, based on teachers' perceptions of problems in the school, (2) teachers' qualifications (that is, their years of teaching experience and attainment of a master's degree), and (3) four organizational characteristics—school and class sizes, normative cohesion, and teachers' sense of control and influence. As a first step in exploring the relations between these factors and achievement scores, partial correlations of the school-based SASS factors[24] with reading and math assessment scores and with each other are shown in table 1. These partial correlations represent the bivariate relations among the factors, partialing out the three background factors of poverty, language barriers, and race-ethnicity.

Partial correlations reveal the contributions of the school-based factors considered singly to achievement variance. However, they do not capture multivariate relations among the school-based factors. For example, are the negative partial correlations of school and class size with reading assessment scores in

[24] Each factor is defined as the average of the measures indicating the factor shown in figure 3, with measures scaled to equal standard deviations.

Table 1. Partial Correlations of Mean Reading and Mathematics Scores with School-Level Factors in Public Elementary, Middle, and Secondary Schools, Controlling for Background Factors

Factor 1	Factor 2	Elementary (n = 1123)		Middle (n = 496)		Secondary (n = 595)	
		Reading	*Math*	*Reading*	*Math*	*Reading*	*Math*
	Student Achievement in Reading & Mathematics						
School Size		–0.10	–0.09	–0.18**	–0.17	+0.11	–0.06
Class Size		–0.10	–0.02	–0.26*	–0.10	–0.12*	–0.11**
Normative Cohesion		–0.01	+0.01	+0.04	–0.06	–0.05	–0.09
Teachers' Influence		+0.03	+0.05	+0.09	+0.20*	+0.07	+0.23*
Teachers' Influence		–0.02	–0.04	–0.05	–0.10	–0.08	–0.16*
Teachers' Qualifications		+0.04	+0.01	+0.12*	+0.05	+0.08*	+0.11*
	School Climate						
School Size		–0.14*		–0.26*		–0.36*	
Class Size		–0.15		–0.24*		–0.26*	
Normative Cohesion		+0.34*		+0.41*		+0.44*	
Teachers' Influence		+0.18*		+0.20*		+0.26*	
Teachers' Qualifications		+0.00		+0.06		–0.08	
	Teachers' Self-Perceptions of Influence						
School Size		–0.13*		–0.14*		–0.22*	
Class Size		–0.09		–0.03		–0.14*	
Normative Cohesion		+0.35*		+0.27*		+0.25*	
Teachers' Qualifications		–0.10*		–0.06*		–0.06	
	Normative Cohesion						
School Size		+0.02		–0.01		–0.09*	
Class Size		–0.07		–0.03		–0.07	
Teachers' Qualifications		–0.01		–0.04		–0.03	

Table 1. Partial Correlations of Mean Reading and Mathematics Scores with School level Factors in Public Elementary, Middle, and Secondary Schools, Controlling for Background Factors (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers' Qualifications			
School Size		+0.11	+0.10	+0.24
Class Size		+0.06	–0.04	+0.11
	Class Size			
School Size		+0.33*	+0.48*	+0.54*

NOTES: Table entries are partial correlations, partialing out poverty, language, and race-ethnicity factors.

(*) p<.05 based on repeated half-sample standard deviations.

middle schools independent effects, or is one of the correlations a byproduct of the high intercorrelation between school and class size? Multiple regression and structural equation modeling (SEM) provide a more interpretable picture of the interrelations among the factors.

Ordinary least squares (OLS) multiple regression is a second step in the analysis of school-based correlates of achievement. Separate equations can be modeled for reading and mathematics assessment scores and also for school climate, normative cohesion, teachers' influence, class size, and teachers' qualifications. Unlike partial correlations, multiple regression differentiates between "predictor" and "dependent" variables, although the interpretation of equations remains merely that a combination of predictors is correlated with the dependent variable. Estimates of standardized regression coefficients for the equations indicated in figure 2 are shown in table 2.[25]

An important limitation of OLS regression is that while variance in each (intermediate) endogenous factor is partially accounted for by other factors, each factor is assumed to be measured without error in predicting other factors. For survey measures, that assumption is not supportable; and a preferable method of analysis is structural equation modeling (SEM), which takes measurement error into account in estimating the proportions of variance in factors

[25] The same factor definitions are used in tables 1 and 2. Estimates of coefficients for background factors are not shown in table 3 because the coefficients are not germane to the exploration of school-based correlates of achievement

Table 2. OLS Standardized Regression Weights for School-level Factors Associated with Mean Reading and Mathematics Scores in Public Elementary, Middle, and Secondary Schools

Factor 1	**Factor 2**	**Elementary (n = 1123)**		**Middle (n = 496)**		**Secondary (n = 595)**	
	Achievement in Reading & Student Mathematics	*Reading*	*Math*	*Reading*	*Math*	*Reading*	*Math*
School Size		-–0.0	–0.07	–0.04	–0.08	+0.29*	+0.11
Class Size		–0.05	+0.01	–0.16*	–0.04	–0.20*	–0.08*
Normative Cohesion		–0.02	+0.00	–0.00	–0.10	–0.11*	–0.18*
Teachers' Qualifications		–0.00	–0.02	–0.04	–0.06	–0.09	–0.12
Teachers' Influence		+0.01	+0.03	+0.04	+0.17*	+0.08	+0.21*
School Climate		+0.02	–0.01	+0.05	+0.02	+0.16*	+0.15*
r^2		0.49	0.36	0.54	0.50	0.48	0.48
	School Climate						
School Size		–0.09*		–0.15*		–0.23*	
Class Size		–0.06		–0.11		–0.06	
Normative Cohesion		+0.25*		+0.29*		+0.31*	
Teachers' Influence		+0.04		+0.05		+0.08	
r^2		0.51		0.58		0.57	
	Teachers' Self-Perceptions of Influence						
School Size		–0.15*		–0.15*		–0.21*	
Normative Cohesion		+0.36*		+0.26*		+0.22*	
r^2		0.18		0.15		0.20	
	Normative Cohesion						
School Size		+0.02		–0.01		–0.10	
r^2		0.07		0.02		0.03	

Table 2. OLS Standardized Regression Weights for School-level Factors Associated with Mean Reading and Mathematics Scores in Public Elementary, Middle, and Secondary Schools (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers' Qualifications			
School Size		+0.12	+0.11	+0.27
r^2		0.03	0.03	0.09
	Class Size			
School Size		+0.35*	+0.51*	+0.55*
r^2		0.14	0.14	0.38

NOTES: (*) p<.05 based on repeated half-sample standard deviations.

r^2 values include effects of three background factors in addition to factors shown.

accounted for by other factors. Estimates of coefficients for the SEM corresponding to the relations in figure 2 and the specified structural equations are shown in table 3.[26]

If all of the school-based predictors in the OLS regression equations were uncorrelated with each other after background factors were taken into account, then the standardized regression weights displayed in table 2 would, to a first approximation, be the same as the partial correlations displayed in table 1. Furthermore, *if* all of the indicators of each factor in the SEM analysis were perfectly intercorrelated, then the results in table 3 should be approximately the same as those in table 2. (The approximation would not be exact because the SEM equations are estimated jointly, while the OLS equations are estimated separately.) Thus, substantial differences in corresponding coefficients between these tables demonstrate the impact of measurement error and multivariate interactions.

Although SEM is the preferred method of analysis for such analyses because it takes into account the measurement error in predictors, estimation of SEM coefficients requires a large number of degrees of freedom, and the

[26] The corresponding structural equation coefficients for the background factors are not shown in table 3 because they are not germane to the exploration of school-based correlates of achievement. As expected, SEM analyses confirm that all measurement variables are indeed significantly related with the corresponding factors, as indicated in figure 3. See appendix A for results of SEM on the measurement level.

Table 3. SEM Associations of Mean Reading and Mathematics Scores with School-level Factors in Public Elementary, Middle, and Secondary Schools

Factor 1	Factor 2	Elementary (n = 1123)		Middle (n = 496)		Secondary (n = 595)	
		Reading	*Math*	*Reading*	*Math*	*Reading*	*Math*
	Achievement in Reading & Student Mathematics						
School Size		+0.10	+0.02	+0.07	–0.13	+0.50*	+0.18
Class Size		–0.34*	–0.18	–0.37*	–0.11	–0.53*	–0.30
Normative Cohesion		–0.03	–0.00	+0.02	–0.14	+0.05	–0.06
Teachers' Qualifications		–0.02	–0.05	–0.14	–0.17		
Teachers' Influence		+0.03	+0.08	+0.00	+0.18*	–0.09	+0.16
School Climate		–0.30*	–0.30*	–0.06	–0.04	–0.09	–0.14
r^2		0.91	0.66	0.92	0.89	0.89	0.89
	School Climate						
School Size		–0.09		–0.08		–0.33*	
Class Size		–0.12		–0.23		–0.001	
Normative Cohesion		+0.24*		+0.37*		+0.34*	
Teachers' Influence		+0.00		+0.11		+0.00	
r^2		0.66		0.698		0.71	
	Teachers' Self-Perceptions of Influence						
School Size		–0.11		–0.043		–0.21	
Normative Cohesion		+0.39*		+0.35*		+0.30	
r^2		0.21		0.28		0.47	
	Normative Cohesion						
School Size		–0.02		–0.13		–0.19	
r^2		0.10		0.04		0.09	

Table 3. SEM Associations of Mean Reading and Mathematics Scores with School-level Factors in Public Elementary, Middle, and Secondary Schools (continued)

Factor 1	Factor 2	Elementary (n = 1123)	Middle (n = 496)	Secondary (n = 595)
	Teachers’			
	Qualifications			
School Size		+0.16	+0.04	
r^2		0.09	0.11	
School Size	Class Size	+0.54*	+0.75*	+0.70*
r^2		0.37	0.50	0.57
		Statistical Summary Measures		
	GFI (AGFI)	0.95 (0.92)	0.92 (0.86)	0.94 (0.88)
	P2 (d.f.)	547 (122)	473 (122)	400 (93)

NOTES: Entries in table are standardized gamma coefficients.

(*) p<.05 based on repeated half-sample standard deviations.

r^2 values include effects of three background factors in addition to factors shown.

resulting estimates can be sensitive to intercorrelations between the indicators of the various factors. Thus, to assess the robustness of the SEM results, it is useful to compare the results from the three approaches. Where they are in conflict, further study of the fit of the data to the assumptions of the method is warranted.

SEM results must be interpreted carefully, however, because the apparent implied causal direction in the regressions is not necessarily valid. Simple interpretation of the results in table 3 requires that one assume that the arrows in the causal model shown in figure 2 imply causality.[27]Any significantly positive or negative coefficient might also reflect a causal relation in the opposite direction.[28] [29]

[27] The selection of factors for exclusion from particular structural equations was based on trial estimation runs in which the goodness of fit index was improved by their deletion.

[28] Table 3 does not include the coefficients for the measurement equations portrayed in figure 3. These coefficients are presented in appendix A in this paper.

[29] Parameter estimates for a variant of the model in figure 2, in which the positions of school climate and achievement are reversed, are given in appendix B in this paper.

A note is necessary concerning the apparently very high percentages of variance accounted for in the reading assessment scores by the SEM analysis (89 to 91 percent). This figure does not indicate the percentage of variance accounted for in the *observed* assessment scores. Because only a single measure of reading assessment was available for each school, the selection of a percentage for measurement error was somewhat arbitrary. However, estimation of the SEM equations did not yield a feasible solution unless measurement error was included and allowed to be correlated between the mathematics and reading scores. Only by setting the measurement error variance to a narrow range of values could a feasible solution be obtained using SAS PROC CALIS. For example, for elementary schools, the measurement error variances for both math and reading scores were set to 0.23. Thus, the school-level factors were accounting for 89 to 91 percent of a factor that itself constitutes 77 percent of the observed school mean reading score variance.[30] Finally, at the high school level, it was necessary to omit the teacher qualifications equation to obtain a convergent SEM solution.

Simultaneously, interpretation of the results presented in tables 1, 2, and 3 provides information about the school-based correlates of achievement and about the extent to which measurement error and correlated predictors can distort analytical results. Through these analyses, we address the following question: "Are school climate, class size, school size, teachers' perceptions of normative cohesion and sense of control, and teachers' experience and education level statistically significantly related to reading and mathematics achievement, based on the SASS student achievement subfile?"[31]

School Climate

The partial correlations in table 1 indicate that a positive school behavioral climate is a correlate of higher average achievement, but only at the middle (for reading) and secondary levels. The OLS regression results in table 2 confirm this finding, but at none of the levels do the SEM results in table 3 indicate a positive relation.

[30] A separate series of analyses not reported here explicitly defined achievement as a single factor contributing to two observed measures, reading and math school means.

[31] For the purposes of inferring statistical significance, standard errors of the tabulated estimates were estimated by replicating the analyses on 100 random half-samples of states in the database.

So is school climate a correlate of achievement in middle and secondary schools? Should educators expect that improving school climate might contribute to improved test scores? Examination of the correlates of school climate in tables 1, 2, and 3 sheds some light on this issue. The results from all three sets of analyses suggest that school behavior climates are better in schools with high normative cohesion (i.e., where teachers feel that they have common goals and cooperate) and in smaller schools, especially among students in higher grade levels. Although cohesion and climate are correlated with each other, the partial correlation of achievement with cohesion is smaller than its correlation with climate, so in the secondary school OLS regressions, cohesion becomes a moderator, with a negative coefficient, magnifying the positive coefficient for climate. Two factors in the SEM model eliminate the effect, suggesting that it is an artifact of OLS regression. First, the contribution of normative cohesion to school climate is explicitly taken into account; and, second, the "method" factor representing teachers' positive or negative response tendencies is included.

The message from these analyses, based on the assumptions embodied in figure 2, is that steps to increase normative cohesion among staff may go hand-in-hand with improving school climate, but that these factors are not strong correlates of assessment scores when measurement error is taken into account.

Class Size

As predicted by most current research, reduced *class size* is related to higher academic performance, but the significant relations are primarily limited to reading in this study; the relations with math scores are much weaker. For reading, the partial correlations with class size in table 1 are significant at the middle and secondary school levels; the regression coefficients in table 2 at the middle and secondary level are significant; and the SEM coefficients in table 3 are significant at all three grade levels.

It seems clear from these data that the class size factor is a correlate of reading achievement among middle and secondary schools. At the elementary level, on the other hand, the partial correlation coefficients and OLS results do not corroborate the SEM results. This appears to be an instance where the flexibility of the SEM method enables it to uncover a relation that is hidden by the imperfect relations between observed indicators and underlying factors.

Even in the SEM results, however, class size is a stronger correlate of reading achievement in higher grades. Student/teacher ratios, reported class sizes, and teachers' satisfaction with class sizes all tend to be more favorable in high schools with higher reading (or verbal or language) scores than in high schools with lower average scores.

School Size

The class size and school size factors are positively correlated with each other at all three grade levels, as shown in tables 1, 2, and 3; but the patterns of their correlations with achievement are different. At the secondary level, reading scores are higher in schools with larger enrollments, but at the elementary level there is no significant correlation.[32] The negative partial correlation for middle schools, shown in table 1, is shown to be artifactual in tables 2 and 3: higher reading scores are found in smaller middle schools because those schools have smaller class sizes. Across all three grade levels, there is a tendency (shown in table 3) for school size to be more positively (or less negatively) associated with reading scores than with mathematics scores.

The relation between school size and climate is clearer: at all three levels, the partial correlations and OLS regression results indicate better climates in smaller schools; the SEM results are in the same direction, although they are only statistically significant at the secondary level. Whether smaller schools happen to be in neighborhoods and communities where teachers perceive better school climates or whether smaller schools foster better climates is an issue for further study.

Teachers' Perceptions of Normative Cohesion and Sense of Control and Influence

Normative cohesion and the sense of having influence on school policies and control over classroom decisions are positively related to each other at all three grade levels. While the expectation is that the cultural cohesion establishes a stable foundation for performance, the results in tables 1, 2, and 3 do not indicate that either of these factors is a significantly positive correlate of achievement. In fact, there is a negative relation between normative cohesion

[32] The school size measure is the logarithm of total enrollment.

and achievement in OLS regression at the secondary level, but this is an artifact of the close relation between the normative cohesion and school climate composites. When measurement error in these constructs is taken into account, by SEM analyses, the negative relation disappears, as does the positive relation of school climate to achievement.

Middle and secondary schools in which teachers perceive that they have more than average control over classroom practices and influence on school policies tend to be schools in which mathematics scores are higher. Further study is needed to determine whether this phenomenon represents a specific effect on mathematics teachers or a general "reform" factor (i.e., some school administrators, more than others, recognize both the critical need for math skills and the importance of empowering teachers). The relations among normative cohesion, sense of control, and school climate are consistent with Ingersoll's (1996) conclusion that teachers' autonomy and influence "make an important difference for the amount of cooperation or conflict in schools" (p. 171). Relying on SASS 1987–88 data with very similar measures, he adds that this relationship varies by the locus of teachers' control: when locus is fundamentally social (i.e., "selection, maintenance, and transmission of behaviors and norms"; p. 171), rather than concerned with curriculum and instruction (i.e., "selection of textbooks, topics, materials, and teaching techniques"; p. 171), then the association between teachers' lack of power and conflict in school is strongest.

Overall, there is evidence in the SASS student achievement subfile that organizational characteristics of schools are correlates of student achievement. SASS offers an abundance of opportunity to assess organizational characteristics in American schools (see Baker 1996). Hence, while this report focuses on four organizational characteristics, future analyses of these data might concentrate on such organizational features as organizational inertia (as, for example, in regard to personnel tenure); organizational change (as, for example, in regard to reform issues); or autonomy (as both an intra-organizational feature—room for initiatives—and an inter-organizational feature—dominance of the school district over major decisionmaking issues).

Teacher Qualifications

The final potential correlate of achievement at the school level examined in this study is the average of teachers' qualifications, as represented by years of teaching experience and attainment of a master's degree. The SASS student achievement subfile did not strongly support the investigation of this factor at the school level, due both to the high level of within-school variance among the responding teachers (i.e., low reliabilities) and to the low intercorrelation between the two indicators (see appendix A in this paper). At the secondary level, the SEM estimation procedure would not even easily converge when these indicators were included. Therefore, the findings of insignificant relations between this teacher qualification composite and reading and mathematics achievement are not unexpected.

Unlike the factors based on teachers' perceptions (school climate, normative cohesion, and sense of control), which can have a communality based on common perceptions of the school as a workplace environment, teachers' experience and education are naturally highly variable within a school, as new teachers are continually added to replace highly experienced teachers who retire. Investigation of these indicators as correlates of achievement appears to require achievement data at the individual classroom level.

It may be that other teacher qualification factors available in SASS, such as matches of college major to teaching assignments and selectivity of the teachers' undergraduate colleges, have sufficient communality within schools to support school-level analyses. Otherwise, study designs such as NELS:88 and NAEP (when specific teacher questionnaires were matched to specific students' performance) are more appropriate for assessing the correlation between teacher qualifications and achievement.

Conclusions

The objective of this report has been to demonstrate and evaluate the strategy of combining a large-scale national survey of schools (SASS), which lacks measures of student achievement, with school-level assessment data from a large number of individual states. If application of this strategy yields new insights about schools or identifies questions that lead to new avenues of research, then its value is demonstrated. If the substantive findings are empty, the strategy is less attractive.

To demonstrate the strategy, a set of 18 composites of SASS data, including student background information, organizational information, teachers' qualifications, and school climate perceptions, were constructed and merged with school reading and mathematics mean scores. The resulting data were analyzed using correlational, multiple regression, and structural equation model analyses. These analyses only begin to tap the richness of the SASS database: selection of other subsets of the SASS data or other analytical methods could add to the evaluation of the strategy.

Substantive Findings

The clearest result with respect to correlates of achievement is that reading scores are higher in schools with smaller class sizes. This result, obtained from structural equation modeling using both state assessment data and NAEP adjustments for between-state variance in achievement, is consistent across grade levels (see table 3). While there are alternative causal explanations for this finding, such a finding in a large sample of public schools in 20 states is an important corroboration of the controlled research results that indicate that class size makes a difference.

The positive relation between small classes and reading scores was stronger for secondary schools than for elementary schools. In secondary schools, the positive association with reading included both large schools and small classes. The relation between class size and achievement was specific to reading scores; it was much weaker for mathematics.

Substantive findings were not limited to class size. Teachers' perceptions that they had control of classroom practices and influence on school policies tended to be greater in middle schools, and possibly in high schools, in which mathematics scores were higher. The analyses carried out do not provide a causal explanation for this relation, but its statistical significance suggests a potentially fruitful area for both additional studies of SASS measures and controlled research.

Because the data are not longitudinal, causal inferences must be treated much more tentatively than conclusions based on data on the achievement gains of a specified set of students over time. Also, because the data are school means, they cannot address the factors that differentially affect the achievement of different students in the same school. Nevertheless, findings from analyses of the SASS student achievement subfile, based on over 2,000 schools in 20 states, can contribute to the overall educational policy database.

Methodological Findings

The primary conclusion reached in this study is that because the data are readily available, the strategy of matching school-level assessment scores to a national survey is feasible, and not costly, and leads to valid and reliable conclusions about correlates of public school achievement across much of the United States. The additional step of linking the database to State NAEP to capture between-state achievement variation provides additional informational value and is also feasible and not costly. In a separate report (McLaughlin and Drori 1999), a comparison between analyses that include the State NAEP adjustment and simple pooled within-state aggregations indicated that some correlates were stronger (e.g., between class size and achievement) when the State NAEP adjustment was used.

A positive methodological finding was the generalizability of the between-state achievement measures across grade levels. Although state assessment scores were available for grades from 3 to 11, NAEP reading scores for individual states were only available for grade 4. If the ordering of states in reading achievement changed substantially from grade 4 to grades 8 and 11, then the results of overall analyses of middle school and high school data would be diluted by linkage error.

The extension of the NAEP adjustment proved valid, in that the findings for secondary schools are as meaningful as the findings for elementary schools. This conclusion is not surprising, given the very high correlation of State NAEP means in different grades and subjects, but its support in this study may suggest new uses of State NAEP data in conjunction with state assessment data.

A limitation on the validity of aggregating teacher data for school level analyses became apparent in the findings concerning teacher qualifications (average years of teaching experience and percent having a master’s degree). These measures, unlike the teachers’ responses to questions about school policies and school behavioral climate, had very low reliability *as measures of the school*, because there was relatively little systematic between-school variation: most of the variation was between teachers at the same school. This problem was manifest in the low intercorrelation between these measures; as a result, the analytical findings concerning the relations of this teacher qualifications factor were uninterpretable. In fact, the SEM software had difficulty even converging to a solution when this teachers' qualification factor was included in structural equations.

Finally, the procedure of examining ordinary least squares regressions and partial correlation coefficients to understand the results of structural equation model analyses proved valuable, in that it suggested explanations for unusual findings, such as the negative coefficient for normative cohesion in predicting middle school achievement. Although the data are purely correlational, there are logical constraints, such as that school factors probably do not cause differences in student background characteristics in the short term ("white flight" notwithstanding). Interpretation of the results of structural equation modeling in terms of hypothetical path models, such as those shown in figure 2, can lead to fruitful suggestions for avenues of research and policy development.

Future Research

Two broad areas of research stemming from this study appear to be fruitful: development of a measure of a school's achievement gains over time which can be associated with SASS measures and further refinement of the linkage functions between state assessments and NAEP.

Every school addresses the needs of a different student population with different resources, and it is therefore unfair to hold all schools accountable to the same achievement standard. However, a number of states are turning to reform criteria that base decision-making on measures of *gains* in achievement over years. Although SASS cannot easily add longitudinal student growth data, it is certainly feasible to add other years' school-level achievement data to the subfile. Specifically, the addition of 1997–98 reading scores, linked to the 1998 grades four and eight State NAEP reading assessment and CCD data on changing enrollment patterns and resources over the intervening years, would provide the basis for identifying SASS factors (measured in 1994) that are predictive of gains in achievement. For example, one wonders whether staff turnover rates would portend gains, other things equal. Of course, states continue to develop and refine assessment systems, and the state assessment test scores for a school in 1998 may not be equivalent to scores obtained in 1994, so linkage of measures of achievement gains over time to repetitions of State NAEP is an essential requirement for the development of a longitudinal database.

The power of the database for longitudinal analyses can be greatly enhanced with the addition of the next cohort of SASS. If a subsample of schools included in SASS in 1994 are also included in 1999–2000, then using the 2000 State NAEP assessment for adjustment of mathematics scores would enable the matching of longitudinal changes in SASS school-based factors with longitudinal changes in achievement, controlling for longitudinal changes in student background factors.

A second line of research would focus on improving the achievement measures included in the SASS student achievement subfile. The linkages used for the analyses presented in this report were based entirely on the means, standard deviations, and correlations between State NAEP and state assessment school means. The errors in these linkages can be diminished significantly by more detailed analysis of the relations among the scores. In particular, current research for the National Center for Education Statistics has found that linkages to NAEP can be improved by considering nonlinear terms and by including demographic indicators. For example, all state reading assessments are sensitive to racial-ethnic differences, but some are more sensitive than others. Their sensitivities could be matched to NAEP’s measurement of the distribution of racial-ethnic achievement differences by explicitly including that matching factor in the NAEP adjustment step in constructing the SASS school-level achievement score. The result would be increased comparability of within-state variation in the achievement measure across states.

References

Baker, D. P. (1996). Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with Comments on School Reform, Governance, and Finance. In J. Mullen and D. Kasprzyk (Eds.), *The Schools and Staffing Survey: Recommendations for the Future* (NCES 97–596) (pp.19–37). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Bollen, K. A., and Bollen, W. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Boruch, R. F., and Terhanian, G. (1996). So What? The Implications of New Analytic Methods for Designing NCES Surveys. In G. Hoachlander, J. Griffiths, and J. Ralph (Eds.), *From Data to Information—New Directions for the National Center for Education Statistics* (NCES 96–901). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature 24*:1141–1177.

Herriot, R. F., and Firestone, W. A. (1984). Two Images of Schools as Organizations: A Refinement and an Elaboration. *Educational Administration Quarterly 20*(4):41–58.

Ingersoll, R. M. (1996). Teachers' Decision-making Power and School Conflict. S*ociology of Education 69*(2):159–176.

Kaufman, P. (1996). Linking Student Data to SASS: Why, When, How. In J. Mullens and D. Kasprzyk (Eds.), *The Schools and Staffing Survey: Recommendations for the Future* (pp. 53–65). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

McLaughlin, D. H. (forthcoming). *Study of the Linkages of 1996 NAEP and State Mathematics Assessments in Four States* (NCES 2000–461). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

McLaughlin, D. H., and Drori, G. (forthcoming). *School-level Correlates of Academic Achievement: Student Assessment Scores in SASS Public Schools* (NCES 2000–303). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Mullis, I. V. S., Campbell, J. R., and Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States.* (No. 23-ST06). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993). *NAEP 1992 Mathematics Report Card for the Nation and the States* (No. 23-ST02). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1996, June). *Profile of 1994–95 State Assessment Systems and Reported Results*. Washington, DC: Author.

Reese, C. M., Miller, K. E., Mazzeo, J., and Dossey, J. A. (1997*). NAEP 1996 Mathematics Report Card for the Nation and the States* (NCES 97–488). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Roeber, E., Bond, L., and Braskamp, D. (1997*). Annual Survey of State Student Assessment Programs: Fall 1996.* Washington, DC: Council of Chief State School Officers.

Roeber, E., Bond, L., and Connealy, S. (1998*). Annual Survey of State Student Assessment Programs: Fall 1997*. Washington, DC: Council of Chief State School Officers.

Wu, G., Royal, M., and McLaughlin, D. (1997). *Working Paper: Development of a SASS School-level Student Achievement Subfile, Using Existing State Assessment and State NAEP Information* (NCES 97–44). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Appendix A
Results of SEM: Measurement Level

A measurement model is required to link the latent variables in the structural equation model to SASS measures. Four tables provide the information necessary for understanding the latent variables.

Table A displays the indicators of each latent trait and the SASS variables used in its computation. For example, poverty is measured by the ratio of the reported number of free-lunch eligible students (S1655 and S1660) to the total enrollment (S0255), and by the average of teachers' responses to the item asking whether poverty is a slight, moderate, or serious problem at the school (T1165).

Table B gives a reliability estimate for the teacher-based indicators, that is, a measure of the tendency of teachers at the same school to give the same responses. The estimate is one minus the ratio of (a) the sum of within-school variances, divided by the total number of teachers responding, to (b) the variance of school means. Values substantially less than 0.5 (such as average class size, classroom control perceptions, and years of experience) indicate that more of the variance in the indicator is within schools than between schools, and alternative indicators should be given greater weight in the model.

Table C gives intercorrelations between the indicators. For example, the lowest intercorrelations (0.28 and 0.29) are for the average class size with the other two indicators of the class size factor and between average years of teachers' experience and percent of teachers with a master's degree, all at the elementary school level. These intercorrelations need not be substantial for SEM analyses, because the estimation procedure should identify the weighting of the indicators that most effectively accounts for variance in other measures. Nevertheless, values substantially less than 0.50 indicate that most of the variance in the indicators is not in common across the latent trait.

Table D contains the SEM measurement parameter estimates which were obtained simultaneously with the structural equation parameter estimates. For example, the latent poverty variable is set to be in the same units as the free-lunch eligible fraction by presetting its coefficient to 1.0; at the elementary school level, 0.65^2 or 42 percent of the variance in the free-lunch-eligible fraction is attributed to the latent poverty trait.

Table A. School-level Achievement Correlates: Factors, Indicators, and SASS Items Used

Category	Factor	What It Means	How We Measured It Our Indicators	SASS Items Used
Student Background				
	Poverty	Low SES	Ratio of lunch program eligibility, and teachers' identification of a poverty problem	S1655 S1660 (by S0255) T1165
	Race	Racial tension and problems	Ratio of white students, and teachers' identification of a problem of racial tension	S0425 (by S0255) T1170
	English Language proficiency	Low proficiency of English	Ratios identified as having language problems (i.e., ratios of participation in English enhancing programs), ratio of Hispanic students, and teachers' identification of limited English as a problem for students	S1410 S1295 S0415 (by S0255) T1190
Teaching Quality				
	Teacher qualifica-tions	Teachers' degrees and teaching experience	Professional education (i.e., having a master's degree) and years of teaching experience	T0235 T0095 T0100 T0105 T0110
School Climate				
	School atmosphere and problems	Student behavior	Teachers' identification of problems with (set #1) tardiness, dropping out, lack of academic challenge, vandalism, drug abuse, physical conflicts, verbal abuse of teachers, physical attacks on teachers, and teacher absenteeism, and(set #2) student absenteeism, cutting class, apathy, robbery or theft, disrespect of teachers, alcohol abuse, and weapons in school	T1005 T1075 T1080 T1085 T1100 T1105 T1115(5)T1150 T1325 T1330 T1340 T1345

Table A. School-level Achievement Correlates: Factors, Indicators, and SASS Items Used (continued)

Category	Factor	What It Means	How We Measured It Our Indicators	SASS Items Used
Organizational Features				
	School size	Organization size	Number of students enrolled	S0255
	Class size	Resource per student; crowding	Average class size, ratio of enrollment to number of teachers, and teachers' sense of satisfaction with class sizes	S0255 T0830-T0970 S0255 S0910 S0850; T1285
	Teacher Influence	Teachers' control over school policies and classroom arrangements	Teachers' sense of influence over such school matters as setting discipline policy, determining content of inservice programs, hiring, school budget, teacher evaluation, and establishing curriculum; and, teachers' sense of influence over such classroom matters as selecting textbooks and other instructional materials, selecting content, topic, and skills to be taught, selecting teaching technique, evaluating and grading, disciplining students, and determining amount of homework assigned	T1015 T1020 T1025 T1030 T1035 T1040 T1045 T1050 T1055 T1060 T1065 T1070
	Normative cohesion	Clarity of norms	Colleagues share beliefs and values; principal enforces rules and backs up staff; receive support from parents; principal lets staff know expectations; goals and clear; principal knows what he/she wants and communicates to staff; behavior rules are consistently enforced.	T1200 T1225 T1245 T1255 priorities are T1260 T1265 T1295
		Cooperation among school staff	Coop among staff; whether teachers coordinate course with other teachers; sense that administration-staff relations are supportive and encouraging; teachers integrate library/media sources into curriculum	T1205 T1270 T1290 T1310

NOTE: SASS items beginning with "T" are on the teacher questionnaire, and those with "S" are on the school questionnaire.

Table B. Reliability Coefficients for Teacher-based Components of School-level Factors, by School Level (Estimated fraction of sample mean variance that is between schools)

		Elementary	Middle	Secondary	All Schools
Class Size	Average	.31	.37	.44	.42
	Satisfied?	.45	.47	.54	.51
Climate	Climate 1	.70	.74	.77	.82
	Climate 2	.78	.78	.86	.85
Normative	Cooperation	.51	.46	.47	.50
Cohesion	Clear Norms	.54	.55	.61	.60
Teacher Control	Classroom Control	.36	.38	.50	.41
	Influence on School Policies	.61	.62	.62	.63
Teacher Qualifications	Years Experience	.26	.14	.27	.24
	Masters	.41	.39	.44	.41
Poverty	Problem?	.82	.79	.82	.80
Minority Conflicts	Problem?	.73	.79	.84	.80

Table C. Intercorrelations of Components of School-level Factors, by Level

					All
		Elementary	Middle	Secondary	Schools
Class Size Factor					
Average Class Size	Student/ Teacher Ratio	.28	.50	.57	.45
Average Class Size	Class Size Satisfaction	.29	.45	.49	.40
Student/ Teacher Ratio	Class Size Satisfaction	.43	.42	.55	.48
Climate Factor					
Climate 1	Climate 2	.61	.72	.50	.72
Normative Cohesiveness					
Cooperation	Clear Norms	.66	.67	.66	.69
Teacher Control and Influence					
Classroom Control	School Policy Influence	.50	.49	.49	.48
Teacher Qualifications					
Years Experience	Masters	.29	.30	.37	.32
Poverty					
Pct. Free-Lunch Eligible	Poverty a Problem	.55	.51	.47	.49
Racial/Ethnic Minorities					
Pct. Minority	Race Conflicts a Problem	.49	.45	.43	.45
Language Minorities					
Pct. Limited English	Pct in ESL Classes	.80	.75	.70	.77

Table D. School-level Factors: SEM Measurement Model Estimates

Indicator	Elementary	Middle	Secondary
Pct. free lunch eligible	=1.0 (.65) Poverty+ .76e	=1.0 (.66) Poverty+.75e	=1.0 (.61) Poverty+.79e
Teachers see poverty a problem	=1.26 (.81)* Poverty -.34 (-.22)* T_Quest+ .54e	=1.13 (.70)* Poverty -.70 (-.43)* T_Quest+ .57e	=1.38 (.86)* Poverty - .11 (-.07) T_Quest+ .50e
Pct. Limited English Proficient	=1.0 (.87) Language+ .49e	=1.0 (.89) Language+ .46e	=1.0 (.85) Language+ .53e
Pct. in ESL instruction	=1.05 (.90)* Language+.44e	=.94 (.81)* Language+.58e	=.98 (.82)* Language+ .57e
Pct. minority	=1.0 (.81) Race+ .58e	=1.0 (.85) Race+ .52e	=1.0 (.83) Race+ .56e
Teachers see racial tension a problem	=.76 (.63)* Race -.48 (-.30) T_Quest+ .71e	=.62 (.50)* Race -1.01 (-.62)* T_Quest+ .61e	=.64 (.54)* Race -.17 (-.11)* T_Quest+ .84e
Years teaching experience	=1.0 (.57) T.Qualif +.82e	=1.0 (.80) T.Qualif+ .60e	=1.0 (.51) T.Qualif+ .86e
Masters degree	=.90 (.51)* T.Qualif+ .86e	=.46 (.37)* T.Qualif+ .93e	=1.42 (.73)* T.Qualif+ .68e
Perceived classroom control	=1.0 (.58) T.Control+ .82e	=1.0 (.59) T.Control+ .80e	=1.0 (.74) T.Control+ .68e
Perceived policy influence	=1.51 (.87)* T.Control+.49e	=1.39 (.83)* T.Control+ .56e	=.90 (.67)* T.Control+ .75e
School size	=1.0 (1.0) School.Size+ .0e	=1.0 (1.0) School.Size+ .0e	=1.0 (1.0) School.Size+ .0e
Average class size	=1.0 (.39) Class.size+ .92e	=1.0 (.61) Class.size+ .80e	=1.0 (.72) Class.size+ .70e

Table D. School-level Factors: SEM Measurement Model Estimates (continued)

Indicator	Elementary	Middle	Secondary
Student/teacher ratio	=1.78 (.70)* Class.size+.71e	=1.31 (.80)* Class.size+ .60e	=1.12 (.81)* Class.size+ .59e
Teachers see class size as satisfactory	=1.53 (.58)* Class.Size + 1.0 (.61) T_Quest+ .54e	=.87 (.45)* Class.Size + 1.0 (.54) T_Quest+ .71e	=.86 (.63)* Class.Size + 1.0 (.64) T_Quest+ .43e
Clarity of norms	=1.0 (1.0) Norm.Co+ .0e	=1.0 (1.0) Norm.Co+ .0e	=1.0 (1.0) Norm.Co+ .0e
Cooperation	=.65 (.65)* Norm.Co+ .76e	=.66 (.65)* Norm.Co+ .76e	=.65 (.65)* Norm.Co+ .76e
School behavioral climate 1	=1.0 (.94) School.Climate -.33 (.21)* T_Quest+ .28e	=1.0 (.85) School.Climate -.78 (.49)* T_Quest+ .17e	=1.0 (.97) School.Climate -.18 (.11)* T_Quest+ .21e
School behavioral climate 2	=.98 (.92)* School.Climate -.35 (-.23) * T_Quest+ .31e	=.93 (.79)* School.Climate -.79 (-.50) * T_Quest+ .34e	=.91(.91)* School.Climate -.27 (-.18)* T_Quest+ .38e

NOTES: Data in table include: Unstandardized coefficient or factor loading (standardized coefficient or factor loading) significance level, factor name, and magnitude of error term. Indicators with an unstandardized factor loading of 1.0 are the reference indicators. Error terms of .0 were set to this value in order to identify the model.

Appendix B
Factors Associated with School Climate and Achievement in Public Schools, Reversing the Causal Order between These Two Factors: Results from SEM Analyses

A separate SEM analysis was carried out in which the school climate trait was omitted from the equation for achievement and the school achievement trait was included in the equation for school climate. (This analysis was carried out using a single school achievement trait, measured by the two indicators of average reading and mathematics scores.)

Independent Factor	**Dependent Factor**	**Elementary (n = 1124)**	**Middle (n = 496)**	**Secondary (n = 595)**
	Student Achievement			
School Size		+0.10	+0.01	+0.39*
Class Size		–0.25*	–0.27	–0.37*
Normative Cohesion		–0.09*	–0.02	–0.00
Teachers’ Qualifications		–0.01	–0.13	–0.12
Teachers’ Influence		+0.02	+0.02	–0.05
r^2		0.69	0.85	0.81
School Size				
	School Climate	–0.05	–0.06	–0.22*
Class Size		–0.20	–0.26	–0.09
Normative Cohesion		+0.22*	+0.37*	+0.34*
Teachers’ Influence		+0.01	+0.11	+0.02
Student Achievement		–0.29	–0.03	–0.11
r^2		0.68	0.75	0.70
	Statistical Summary Measures			
	GFI (AGFI)	0.95 (0.92)	0.91 (0.85)	0.93 (0.88)
	P2 (d.f.)	606 (130)	552 (130)	506 (130)

NOTES: Entries in table are standardized gamma coefficients.

(*) p<.05 based on repeated half-sample standard deviations. The same factors and equations were included as represented in table 6, except for the reversal of achievement and school climate.

r^2 values include effects of three background factors in addition to factors shown.

Response: Opportunities for Design Changes

Valerie E. Lee
University of Michigan

I have had much experience in the statistical analysis of NCES data sets, particularly those with a longitudinal design. I feel very grateful to NCES for collecting these excellent data and making them available to researchers. Imagine the cost if we had to collect such data ourselves! However, in conducting many studies, it is not unusual for researchers to discover many difficulties—and to want to change future data collections to avoid such difficulties.

Thus, I am using this conference, and my participation here, as a public opportunity to encourage NCES to consider some changes in their data collection designs. I believe that the implementation of these suggested changes would help researchers address policy-relevant questions about education more accurately and reliably. I use my comments about the papers in this session as a "launching pad" for making these suggestions.[1]

Comments on the Brewer and Goldhaber Paper

The Brewer and Goldhaber paper, "Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88," is a solid, even classic, example of how economists conduct educational production function research. The dependent variable these researchers explore is a measure of learning, which they conceptualize as the change in students' achievement over a two-year period—the first two years of high school. The authors focus on achievement gains in the four subjects that were tested in NELS. I agree with the approach that uses gains in achievement in individual subjects as measures of learning.

[1] The opinions expressed in this paper are those of the author. Address any comments to Professor Valerie Lee, School of Education, University of Michigan, 610 East University Avenue, Room 4220, Ann Arbor, MI 48109.

The Research Focus

Brewer and Goldhaber's research focuses on estimating achievement gains as functions of classroom conditions and qualifications of teachers. Their analyses take typical control variables into account. The researchers capitalize on the NELS data collection strategy, in which data collected from two of each student's teachers may be matched directly to each student in the subject in which the teacher has taught the student in his or her tenth grade year. In their analyses, these researchers use a single unit of analysis—the student—and they use ordinary least squares (OLS) regression as their major analytic method. From my vantage point, the research described by Brewer and Goldhaber is characterized by a few methodological difficulties, some of which the authors could not have avoided. As stated, I use these difficulties to highlight some shortcomings of NELS data collection designs. I recognize that economists, such as these authors, may have several differences of opinions with sociologists of education like me. Nonetheless, let me spell out two major difficulties.

Design Difficulty Number One

By intent, the NELS data are not designed so that several students are "nested" within each teacher. Moreover, even if several NELS students do have the same teacher for the same subject at tenth grade (i.e., they are matched to the same teacher in this data set), the students are not necessarily in the same classes even when they have the same teacher. To me, the most logical mechanism through which a teacher's qualifications might be linked to a student's learning would be that the teacher's qualifications would somehow influence how he or she teaches, i.e., through instruction. To oversimply this mechanism, let us assume that instruction is to be broken down into two components. One component is related to the content the teacher imparts. Content would take place at the level of the individual class, and teachers would vary their content depending on the class they were teaching (and the students in the class). The second component is the teacher's expertise. This component is one that the teacher would "carry" with him or her regardless of the class being taught; it might be constant across the students (and classes) taught. In the analyses described by Brewer and Goldhaber, and fundamental to the structure of the NELS data, we must focus only on the second component of qualifications: subject matter expertise. The first component, pedagogical or instructional variation by class, must be ignored.

Therefore, with NELS, we are unable to differentiate between the sort of instruction teachers actually use and the expertise they bring to the subject. With this data structure, where there are very few teachers per student and not all the students taught by a single NELS teacher are actually in the same class, we are unable to either theoretically specify or empirically differentiate whether teachers' qualifications—a major independent construct in this research—actually influence student learning through what material is taught, i.e., content, or how material is taught, i.e., expertise level. To me, education is fundamentally about such things. However, the structure of NELS data limits researchers' ability to tease out the differences between content expertise and pedagogical expertise within the construct of "teachers' qualifications." I do not think there is a single thing Brewer and Goldhaber could do to solve this vexing problem. I suggest, however, that they could have recognized it in their paper.

Solution to Design Difficulty Number One

I believe there is a solution to this problem, and it has to do with changing the basic data collection design. If more students were sampled per school, even at the cost of sampling fewer schools, this important problem would be solved. With such a design, we could recognize that students are in fact "nested" both within classrooms as well as within teachers, and that teachers' basic instruction *does* change across the classes they teach. This is not the first time I have made this suggestion to staff at the National Center for Education Statistics. However, I do not want to lose this golden opportunity to say it once again—that NCES should actually include multiple students in particular classrooms as part of their sampling strategy. In this way, we researchers would be able to investigate the nested nature of schooling at three levels: students, nested in classrooms, which are in turn nested in schools. That is, we would be able to tease out these important levels of variability in the nature of schooling.

Design Difficulty Number Two

The achievement outcome that these researchers use—change over time on the same test—has been used by many of us very frequently. As mentioned, I see this as a measure of learning, specifically growth in achievement in particular subject areas between the eighth and the tenth grades. This way of measuring learning capitalizes on the value of longitudinal research, something that is surely not lost on the participants in this conference. It is hard to think of anyone who would not laud this feature of NCES databases as crucially valuable. Despite the value of longitudinal data, there is a basic design flaw endemic in this type of research. Brewer and Goldhaber link learning between eighth and tenth grade—a two-year period—to the qualifications of the teachers to whom the students were exposed in tenth grade only. The obvious problem remains: What happened in the ninth grade? At one level, one could conclude from this paper that Brewer and Goldhaber's findings probably represent an *underestimate* of relationships that might actually be there, simply because the authors were looking at only part of the students' educational experiences over the two-year time period captured by the achievement gain. Further, high schools students could actually have been in these teachers' classes for only a single semester.

Solution to Design Difficulty Number Two

The obvious solution is to collect achievement and instructional information every year. That is, if we want to link instruction to learning using NCES data, we need to match the data collection period to the instructional period; such a match represents a basic issue of construct validity. If we wish to investigate links between the character of instruction and students' learning as a result of that instruction—a fairly basic question in educational research—then we need an appropriate data structure to be able to do so. The NELS structure, with biennial data collection, does not allow us to be able to do so. Unfortunately, this design means that most of the research conducted with NELS data does not really look at instructional effects. One possibility for researchers is to capitalize on the fact that many school districts in the United States test all students annually. For example, I have conducted a couple of studies using an excellent data set collected in Chicago, where the district tests every student every year. These district data are linked to data collected through periodic surveys conducted by the Consortium for Chicago School Research. The excellent cooperation between the Consortium and the Chicago Public Schools has allowed researchers to link survey data to annual test data. As a result, analysts may measure change in achievement over one year on the same test, and these achievement gains may be linked to survey data from students and teachers about students' educational experiences during the same year. The Consortium didn't have to collect achievement data themselves; it was a major political accomplishment to be able to combine data of this type. I suggest that if this has already been accomplished in Chicago, perhaps it can happen at the national level. Rather than constantly collecting new data, we should at least think about how to capitalize on existing test data (and equate scores across different tests), as there is surely enough testing going on in the world.

Summary of Comments on the Brewer and Goldhaber Paper

I agree with these authors on three issues and disagree with them on three other issues. First, I agree with the authors that NCES decisions to collect data from teachers that can actually be linked to students are a real advance over previous NCES data collection efforts in NELS. I would like to think of that advance as the first stage in a data collection design that could actually be vastly improved. The first stage is surely important. Second, these authors and I also agree on the need for a better sampling design. In their paper, they argue that having more students sampled per school, and possibly sampling by classroom, would be useful. Third, I also agree with them about the potential usefulness of being able to link data that are collected from SASS with NELS (SASS is a cross-sectional NCES data collection effort.) SASS has information about so many schools and teachers. At present, we cannot link these two datasets, because each has sampled different sets of schools. In the future, it would be useful to have the longitudinal data collections occur in schools that are also part of SASS.

I disagree with Brewer and Goldhaber on three other issues. The first is their use of a single unit of analysis and a methodology restricted to single-unit analyses: OLS regression. That would be acceptable if we knew that a very low proportion of variability in the outcome occurred anywhere but between students. However, other researchers have shown that with the NELS achievement tests, perhaps 25 to 30 percent of the total variability in these test scores lies systematically *between schools*. The authors ignored this. Therefore, they assumed that there was no variability between schools. This is a serious oversight to me. I teach courses in hierarchical linear modeling (HLM), a methodology that is meant to address the multilevel nature of educational data. I feel a strong need to state what may seem obvious: multilevel questions call for multilevel methods.

A third issue about which I disagree with Brewer and Goldhaber is their lack of attention, in their paper and in their brief summary statement in their presentation, to comparisons between public and private schools in other research they have done. I found the simple mention of this type of comparison, without giving much detail about it, to be quite problematic. I have, with several colleagues, conducted a substantial amount of research on cross-sector comparisons. I would like to have seen what they did, and I would like to have been able to comment on it. Presenting findings in a very summary form, without making them transparent enough so that readers can figure out the analyses from which those summary findings have come, presents a problem. In papers like this one, authors should provide sufficient detail for the benefit of interested readers. I suggest that otherwise such issues don't get raised at all. Some questions that come to mind include: "There are different types of private schools, but did the authors take that into account?" and "What about the social distribution of achievement—a measure of social equity in schools?" These are questions I have examined in some detail in research comparing public and private schools. Other researchers have done this type of research as well.

Comments on the McLaughlin and Drori Paper

The second paper in this session, "School-level Correlates of Reading and Mathematics Achievement in Public Schools," is the work of McLaughlin and Drori. As I see it, the major purpose of the research described in this paper is to demonstrate to the research community that it is possible to merge data from different sources, collected for different designs and for different purposes, into a single data source.

Two Ways to View This Pape*r*

Although it was not exactly clear how McLaughlin and Drori would like readers to view their paper, there are two rather different ways that readers could actually make use of this work. In one approach, readers would regard the analyses heuristically. Viewed through this lens, the authors would like us to see their work as an example of what could be done, i.e., how different NCES data sets could be linked. A major research question taking the heuristic approach might be as follows: "Can researchers really make use of data that are merged in this way?" Using a second lens, readers would consider the paper's substantive analyses and findings. This lens would result in research questions of the following type: "In which types of school do students achieve at higher levels?" Personally, I am more comfortable with the heuristic approach. Given the methodological and conceptual difficulties, I am cautious about the validity of findings and any substantive conclusions drawn from them.

A Problem of External Validity

Both in the presentation and in the paper on which it was based, Don McLaughlin provided much detail about the sample and the methodology used for linkage. Therefore, those details do not need repeating. There are, however, a few issues I would like to mention. One issue is the potentially differential weighting of data by state. For example, the number of sampled schools per state ranged widely: between 50 and 335. Moreover, these numbers are not always proportional to the population. Given that the analyses are at the school level, that kind of between-state variation leads to weighting some states up considerably, while weighting down others. As I see it, this differential weighting arises from two sources: (1) the differential correlations of some state-level information with data from NAEP and (2) the fact that the researchers have a substantial amount of information from some states and very little from other states. Without taking the systematic differences in data quality by state into account, it is risky to generalize these results even to the 20 states for which data are available. Let us consider the outcome variable: school average achievement. The authors first created within-state school averages of students' standardized test scores in several subjects, and they then introduced between-state adjustments using information from NAEP. As McLaughlin mentioned, the authors also weighted these scores by the degree to which the state assessments were correlated with NAEP. The weighting varied considerably—between 0.37 and 0.86. This resulted in schools' average achievement scores in some states (e.g., Kentucky) being weighted down quite substantially. I am not trying to dwell on the specifics here. Rather, I am raising an issue of external validity. For whom, really, are the results to be generalized?

Unit of Analysis, Revisited

In discussing the Brewer and Goldhaber paper, I raised the issue of the appropriate unit of analysis. In that paper, the researcher used a single-level analysis at the level of students to attempt to assess how teacher quality influences students' learning. I suggested that such issues are more appropriately addressed with multilevel methods such as HLM. This same issue—deciding on the appropriate unit (or units) of analysis—is also relevant for the McLaughlin and Drori paper. Whereas Brewer and Goldhaber assumed in their analysis that all of the variance in achievement was between students (although we know that 25 to 30 percent of the variance in achievement lies systematically between schools), McLaughlin and Drori have, by conducting school-level analyses, assumed that all of the variance lies between schools, i.e., they have ignored the large proportion of variance that lies between students in the same schools. In more technical terms, the first set of authors assumed an intra-class correlation in the dependent variable of 0, and the second set of authors have assumed that the intra-class correlation in school achievement is 1. Neither assumption is correct. Brewer and Goldhaber, as I mentioned, had systematically ignored the 25 to 30 percent of the variance in test scores that was between schools. On the other hand, McLaughlin and Drori have ignored the 70 to 75 percent of the variability that is between students within schools.

A Few Other Issues

A few other issues about the McLaughlin and Drori paper relate to the difficulty in drawing substantive conclusions from their results. One issue relates to aggregation bias. These authors have used only school-level variables, many of which they have created through aggregating student-level data. Quite simply, aggregated variables do not typically have the same meaning as the individual variables from which they are created. For example, school average SES may measure the types of students who attend a school, but it also measures resource availability in the school. Yet it doesn't measure the social homogeneity of the student body (which the individual-level measure would capture). Another relevant issue is model specification. The original model with which the authors introduced the paper was very broad. Moreover, there were multiple variables introduced to operationalize each of the constructs they wished to examine. In my opinion, the original model was too broad. The fact that a large number of variables were introduced into the analysis, but only a few "survived," suggested an approach typified by the following: "Well, let's see what counts?" I prefer causal models that are derived theoretically, that are more focused in scope. I admit that each of the constructs was discussed in a theoretical context in the early part of the paper. Still, the model was too broad.

A final issue I would like to discuss relates to drawing causal conclusions from cross-sectional data. The authors were, in fact, quite careful throughout the paper to introduce multiple disclaimers about not conducting causal analysis. However, by its very nature, the type of analysis conducted here is causal. The authors selected a dependent variable (average school achievement) and a large number of independent variables. I believe that we are in fact conducting causal analyses whenever we employ typical general linear model methods, regardless of the disclaimers we introduce. The findings, moreover, were not always stable. Some independent variables were statistically associated with school average achievement using one analysis technique, e.g., structural equation modeling, but not with another technique, e.g., OLS regression. Which set of results should we believe?

A Contradictory Conclusion

The authors of these papers surely did not expect that the two papers would be compared. However, comparison seems almost inevitable, since I've been asked to comment on both of them. Again, the issue is as follows: "Which results do we believe?" Both sets of authors investigated whether teachers' qualifications influence student learning. In their paper, Brewer and Goldhaber concluded that teachers' qualifications count. Although McLaughlin didn't mention it in his oral presentation, in his paper he reached the conclusion that teacher qualifications do not count. I realize that the two papers used different data, used different approaches, and examined the question at different units of analysis. Yet both used data that are at least meant to be nationally representative, and both addressed the same overarching research question. As readers interested in drawing substantive policy conclusions from quantitative research, what are we supposed to learn from two papers with divergent conclusions?

Final Words

Let me end by briefly summarizing the issues I wish to emphasize. Both papers have posed what are essentially multilevel questions, but neither has used multilevel methods. Both are essentially posing questions about school effects: "How do characteristics of schools and the classrooms in them influence the learning of students educated in those schools and classrooms?" Such questions require the use of multilevel methods. A second major issue is how we should measure the effects of instruction on learning. I have offered some suggestions here about more appropriate designs for data collection. If we want to link data on teachers and teaching to student achievement and learning, we need data on more students in each school, probably a sampling design where students are systematically nested within teachers and classrooms. This is important. In my opinion, the "action" in education really is right there in the classroom. So let us get good data so that we can actually explore education at this level. I do not mean to be ungrateful to NCES; they have been helpful to me. Nonetheless, in the spirit of this conference, I think improvements can be made, our methods can be done better, so that is what I am saying. Let us do them better.

Section IV.
Policy Perspectives and Concluding Commentary

Assessment Trends in a Contemporary Policy Context 249
Marshall S. Smith

Educational Research and Educational Policy: An Historical Perspective 279
Christopher Jencks

Improving Research and Data Collection on Student Achievement 299
Brenda J. Turnbull, David W. Grissmer, and J. Michael Ross

Assessment Trends in a Contemporary Policy Context

Marshall S. Smith
Under Secretary of Education
Acting Deputy Secretary of Education
U.S. Department of Education
Washington, DC

For the last 30 years, the “gap” between the scores of African American and white students on standardized tests of reading and mathematics has been a thorny and controversial issue. Efforts to understand—and reduce—the gap have highlighted the challenge of simultaneously making progress toward the two co-equal guiding goals of American public education: educational excellence and educational equality.

In terms of both methodology and research perspectives, the papers presented in this seminar book make thoughtful contributions to our understanding of the achievement gap. These papers, along with those in the Jencks and Phillips (1998) volume, enrich our understanding of policies and programs that will raise overall student achievement while, at the same time, helping to close the achievement gap between black and white students.

Putting such policies in place is a primary focus of the legislative proposal that President Clinton sent to Congress for the reauthorization of the Elementary and Secondary Education Act (ESEA), a task that the Congress is undertaking during the 1999–2000 session.[1] The proposal is based on powerful evidence that *all* of our children can learn to far more challenging standards than we have hitherto understood, as well as a belief that access to high-quality

[1] At the time of the seminar, I was serving as Under Secretary of Education and Acting Deputy Secretary of Education at the U.S. Department of Education. In February 2000, I moved to Stanford University as a Professor of Education, having previously served on the faculties of Harvard University and the University of Wisconsin and also as Dean of the Graduate School of Education at Stanford University.

education resources is a fundamental right for all children.[2] Our work on the ESEA proposal has been, and will continue to be, informed by efforts to identify and understand interventions that promote both educational excellence and educational equality.

The search for such interventions is not new. During the late 1960s and early 1970s, I was one of a small group of researchers who spent months examining hundreds of computer printout pages, searching for explanations of black-white test score differences. Typically, the data on those printouts were from the "Equality of Educational Opportunity Survey" (Coleman et al. 1966). Those cross-sectional data were not as powerful for our purposes as the longitudinal data now available in the High School and Beyond (HS&B) survey and its successor longitudinal surveys, such as the National Education Longitudinal Study (NELS), or as the trend data now available from the National Assessment of Educational Progress (NAEP). Researchers have obtained a clearer picture of trends in achievement and of the nature of the achievement gap as we have gradually accumulated more national and state NAEP assessment data over time. The clearer picture and the trend data are invaluable in enlightening our efforts; still, we face many challenges in both understanding and reducing the gap. These challenges are, of course, the reasons for convening this seminar.

In this paper, I will first examine some plausible hypotheses for the narrowing of the gap that occurred between 1971 and 1988. Second, I will consider reasons why the gap did not continue reducing after 1988 and will examine NAEP trend data for 1990–1996, with a special focus on 1990. Third, I will propose the argument that reforms in the education system from 1992–1999 in many states are leading, at present, to improvements in teaching and learning and will lead eventually to improvements in student achievement. Moreover, I believe that, if well implemented, these changes will lead to further reductions in the gap. Then I will review 1990–1998 NAEP data from assessments that are more extensive than the NAEP trend assessments. These new data indicate that the reforms may already be supporting overall increases in test scores for both African Americans and whites. Finally, I will comment on the use of experimental methodology in education research and raise some questions about how to measure student achievement.

[2] See O'Day and Smith (1993) for a discussion of the impact of standards-based reforms on equality of educational opportunity.

The Achievement Gap Narrows: 1971–1988

In 1990, Jennifer O'Day and I (1991a) examined the achievement results on NAEP trend assessments administered from 1971 through 1988. Over that 17-year time period, the NAEP trend data show that the gap in test scores between African American and white students narrowed substantially for both reading and mathematics achievement. Our analysis considered a number of plausible hypotheses for this reduction in the black-white achievement gap between 1971 and 1988. What we saw—and tried to explain—was a pattern of consistent and substantial increases in African American achievement and almost no change in white scores. The increases in black scores and the relative stability of white scores, taken together, produced a reduction in the gap between the reading scores of black and white students—in less than two decades—of approximately 33 percent for 9-year-olds and over 50 percent for 13- and 17-year-olds. For math, the reductions ranged between 25 and 40 percent over the three age levels. Put in another metric, on the reading assessments administered in 1971 and 1988, the original four grade-level difference between 17-year-old African American and white students in reading narrowed by well over two grade levels.

There was, however, a cloud on this otherwise bright horizon. In our paper, O'Day and I (1991a) suggested that the factors that we believed contributed to the reduction in the gap in the past might have run their course and that other factors were appearing that could lead to a widening of the gap. Two years later, in the context of a paper on educational equality and standards-based reform, O'Day and I (1993) returned to consider the black-white test score gap. To our dismay we found, in the 1990 test score data, preliminary evidence on reading achievement for our hunch. I will address this issue later in this paper.

Now, let us examine the trend data where the reductions in the gap became clear. Table 1 shows the NAEP long-term trend reading scores for whites and blacks at three age levels for test administrations from 1971 to 1996. Table 2 displays the NAEP long-term trend mathematics scores for the same three age levels for test administrations for a shorter period of time, from 1973 to 1996. [3]

[3] Tables 1 and 2 in this paper are taken from *NAEP Trends in Academic Progress, 1996*, rather than from Smith and O'Day (1991a).

Table 1. NAEP Long-term Trend Reading Scores, 1971–1996: National, Public and Private School Students

	AGE 9				AGE 13				AGE 17			
Cohort	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff
1954									**1971**	291.4	238.7	52.7
1958					**1971**	260.9	222.4	38.5	**1975**	293.0	240.6	52.4
1962-63	**1971**	214.0	170.1	43.9	**1975**	262.1	225.7	36.4	**1980**	292.8	243.1	49.7
1966-67	**1975**	216.6	181.2	35.4	**1980**	264.4	232.8	31.6	**1984**	295.3	263.6	31.7
1971	**1980**	221.3	189.3	32.0	**1984**	262.5	236.3	26.2	**1988**	294.7	274.4	20.3
1973									**1990**	296.6	267.3	29.3
1975	**1984**	218.2	185.7	32.5	**1988**	261.3	242.9	18.4	**1992**	297.4	260.6	36.8
1977					**1990**	262.3	241.5	20.8	**1994**	295.7	266.2	29.5
1979	**1988**	217.7	188.5	29.2	**1992**	266.4	237.6	28.8	**1996**	294.4	265.4	29.0
1981	**1990**	217.0	181.8	35.2	**1994**	265.1	234.3	30.8				
1983	**1992**	217.9	184.5	33.4	**1996**	267.0	235.6	31.4				
1985	**1994**	218.0	185.4	32.6								
1987	**1996**	219.9	190.0	29.9								

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: *NAEP Trends in Academic Progress, 1996*. National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Table 2. NAEP Long-term Trend Mathematics Scores, 1973–1996: National, Public and Private School Students

	AGE 9				AGE 13				AGE 17			
Cohort	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff	Test Year	White Scale Scores	Black Scale Scores	White Black Diff
1961					**1973**	274.0	228.0	46.0	**1978**	305.9	268.4	37.5
1965	**1973**	225.0	190.0	35.0	**1978**	271.6	229.6	42.0	**1982**	303.7	271.8	31.9
1969	**1978**	224.1	192.4	31.7	**1982**	274.4	240.4	34.0	**1986**	307.5	278.6	28.9
1973	**1982**	224.0	194.9	29.1	**1986**	273.6	249.2	24.4	**1990**	309.5	288.5	21.0
1975									**1992**	311.9	285.8	26.1
	1986	226.9	201.6	25.3	**1990**	276.3	249.1	27.2	**1994**	312.3	285.5	26.8
1979					**1992**	278.9	250.2	28.7	**1996**	313.4	286.4	27.0
1981	**1990**	235.2	208.4	26.8	**1994**	280.8	251.5	29.3				
1983	**1992**	235.1	208.0	27.1	**1996**	281.2	252.1	29.1				
1985	**1994**	236.8	212.1	24.7								
1987	**1996**	236.9	211.6	25.3								

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: *NAEP Trends in Academic Progress, 1996*. National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

For both tables 1 and 2, a step-wise dashed line separates the data from before and after 1990; therefore, the reduction in the achievement gap that occurred between 1971 and 1988 can be revealed more clearly; and the “cloud on the horizon” represented by the 1990 data can be seen as it first appeared. One way of interpreting these tables is to think of 10 scale points as being roughly equal to a grade level. (A more precise estimate is that a grade level of growth between 9- and 13-year-olds is equivalent to about 12 scale points, while between 13- and 17-year-olds a grade level is about 8–10 points.) A brief review of these findings indicates *extraordinary* changes, as noted here:

- In reading, over a period of 17 years from 1971 to 1988, the scores for black students tested at 9 years of age increased by 18 scale points. Scores for black 13-year-olds increased by 20 points. Both increases represent an improvement of roughly 1.5 grade levels. During the same period, the scores for white students increased only slightly. Thus, the gap between black and white scores was reduced by almost 1.5 grade levels. Over the same years, scores for 17-year-old black students increased by substantially more than 3 grade levels, and the gap reduced from 53 to 20 points.
- In math, over a period of only 13 years from 1973 to 1986, the scores for 9-year-old black students gained 1.5 grade levels, and the gap narrowed by 0.5 grade levels. The 13-year-old black students improved by almost 22 points, over 1.7 grade levels, and the gap narrowed by almost 2 grade levels. Black 17-year-olds increased their scores by approximately 1 grade level, and the gap narrowed by 1.5 grade levels.[4]

Another way of looking at these data is to consider what happens to a group, or birth cohort, of students who are born in a certain year, in other words, to understand the context of their schooling. Fortunately, the NAEP trend data give us the opportunity to review data for several birth cohorts of students. For example, we have data for students born in 1971 for all three age levels that were administered the NAEP reading assessments: the 9-year-olds took the test in 1980, the 13-year-olds in 1984, and the 17-year-olds in 1988. Looking at these data by cohort gives us the chance to examine whether changes in the test scores and in the gap occurred in the early or later years of school-

[4] See also the discussion in Smith and O’Day (1991a).

ing. Table 1 displays reading scores by cohort.[5] Looking above the dashed line, what do we see in table 1 for data collected prior to 1990? Two points seem important. First, following the progress of the cohorts, we see the gap clearly closing. But we also see that the changes did not happen only during the years from 1971 to 1988. The changes stretch from 1962 to 1988. The 9-year-olds tested in 1971 were born in 1962. This reminds us that, when we explore possible ideas about the causes of changes in scores, we need to think about the time periods the students lived through as they were growing up.

Second, these data on reading achievement give us insights about the effects of schooling on students. The data for the youngest group, the 9-year-olds, may be viewed as reflecting a combination of student background factors (family status, preschool attendance, etc.) and the early years of schooling (e.g., grades K–4). The reading scores of 9-year-old black students from the 1962 cohort to the 1971 cohort show an increase of more than 1.5 grade levels. Recall that these 9-year-olds were tested at the beginning and the end of the 1970s, a period that included the maturing of both Head Start and Title I. The scores for black 9-year-olds in the latter two cohorts (1975 and 1979) showed no increases. White 9-year-olds showed a slow increase over the five cohorts. This resulted in a reduction in the size of the gap, suggesting that, during the early years of this period of time, there were substantial relative improvements in preschool and early elementary school opportunities for black students. And the test score increases from this early period were sustained.

But what happens during the periods from 9- to 13- and from 13- to 17-years-old across the cohorts? Of particular interest is that black students generally “grew” more than whites during the 9- to 13-year age period on the reading test. In this age period, for each of the cohorts, the growth in reading achievement for blacks (about 53 points on the average) exceeded that of whites (about 45 points on the average) with a striking difference of 14 points in the 1975 cohort. The picture is more complicated when the 13- to 17-year-old growth estimates are examined. One fact that stands out is that, in these years, the growth for both groups is, on average, only about 30 scale points or about 60 percent of the average growth for the 9- to 13-year-olds. Thus, there is greater growth on these tests for young adolescents than for older adolescents.

[5] Note that for some of the cohorts we took the liberty of using immediately adjacent years. For example, in the 1962–63 combined cohort, both the 9- and the 13-year-olds who were given the test were born in 1962, while the 17-year-olds were born in 1963.

Another feature of these data is that, in the 1958 and 1962 cohorts, the growth for whites was far greater than the growth for blacks. This changed dramatically to no difference in growth in the 1966 and the 1971 cohorts, perhaps indicating some equalization of opportunity in secondary schools.

Plausible Explanations for These Changes

The essay O'Day and I (1991a) wrote was not the end product of a comprehensive research project, but a more general academic overview of education reform. Rather than a detailed statistical analysis, we set out to write a provocative analysis that reviewed broad trends in education policy over the years and related them to changes in test scores.[6] O'Day and I were writing almost 25 years after *Equality of Educational Opportunity* (Coleman et al. 1966) was published, so we framed our text according to that report. Coleman and his colleagues had separated their "explanatory" variables into four clusters: (1) *background and home environment*; (2) *school context* including the social class and racial composition of the school; (3) *teacher characteristics*; and (4) *school resources* including curriculum, libraries, and other items. For our analysis, we combined the third and fourth clusters into one large school resource category, leaving us with three clusters.

Our approach was simple. We reviewed the status of and trends in a number of key explanatory variables within the three clusters and then looked at the scores in the NAEP trend data; we then attempted to relate the two sets of trends to each other. Naturally we were interested in explanatory variables that showed a relationship to student achievement and had a track record in the research literature. We were particularly interested in variables that showed a positive trend or change for blacks, positive in the sense that they indicated that black achievement might rise due to changes in the environment or conditions that were measured by the variable. We wanted to explain the substantial increases in black achievement that the NAEP trend data revealed. To refine our search even more, we also were interested in the same variables showing only a neutral or negative change for whites, because we wanted to be able to understand why white achievement did not increase. Here I will simply summarize a few of our central conclusions.

[6] Seven years later David Grissmer and his colleagues developed what now stands as the definitive quantitative work on NAEP data and the test score gap (Grissmer, Flanagan, and Williamson 1998; Grissmer et al. 1998; Grissmer and Flanagan 1998).

Background Factors

Consider again the period of time from 1962 to 1988. This era spans the period from the beginning of the “Great Society” to the end of the Reagan Presidency. The early years of this period, in particular, witnessed substantial strides in economic well-being for families of African American children, with a smaller improvement for white students. The percentage of black children living in families below the poverty line fell from 65 percent in 1960 to 42 percent in 1980 (Smith and O’Day 1991a). Poverty rates also diminished for families of white children; but a substantially smaller percentage of whites were affected, at least in part because the white percentage in poverty was (and still is) much lower. Given the consistent relationship between poverty and achievement, it is reasonable to identify these changes in economic conditions as contributing to improvements in the achievement of both groups, but to a larger extent for black students.

Another related development occurred between 1970 and 1988. The percentage of mothers of black elementary school children who had completed 12 or more years of schooling nearly doubled, moving from 36 percent to 69 percent, while changing only slightly for white mothers (Smith and O’Day 1991a). Studies of academic achievement consistently find the educational attainment of mothers to have a clear and strong relationship to their children’s educational achievement.

Finally, preschool attendance is another factor that is moderately related to student achievement, particularly in the early grades. From 1960 to 1980, preschool attendance increased substantially for low-income children (Smith and O’Day 1991a). Because there was a higher proportion of low-income black families, the increase in preschool attendance gave proportionally greater educational opportunities to black children.

Yet the demographic picture was not entirely positive. During the same period, a few other factors related to poverty and to achievement increased substantially more for blacks than for whites, such as the percentage of single-parent families. Other factors held level for blacks while declining for whites, such as the number of low-birthweight babies (Smith and O’Day 1991a).

Social Context

Two important factors that influenced African American children positively from 1962 to 1988 were rural-to-urban migration and desegregation.

From 1960 through the early 1980s, large numbers of African Americans from the rural South migrated to urban communities in the South and the North. For those who remained in the South, the rapid development of metropolitan areas and widespread economic recovery during the 1970s and 1980s were significant forces of change that created important opportunities related to improved academic achievement (Smith and O'Day 1991a).

Second, from 1969–1972 most Southern states experienced large-scale desegregation of their schools. Desegregation had powerful effects on the racial and social class composition of schools, improving conditions that are generally positively related to student achievement for blacks and whites, though more so for low-income students. And desegregation affected more than just the composition of the schools. In the South, particularly, there were dramatic improvements in the characteristics of the new schools that black students attended, as they moved from segregated all-black schools into newly desegregated, formerly all-white schools. In sum, educational opportunities increased, especially for black students in the South (Schofield 1991).

Table 3 reflects some of these changes. It shows that the reading gains for blacks across the NAEP assessments from 1971 to 1988 were much stronger in the South and border states than in the Northeast and Midwest states, reflecting the forces of rural-to-urban migration, desegregation, and economic recovery. The scores for white students in the South and border states also increased more than in the Northeast and Midwest, especially for 9- and 13-year olds, but their gains were nowhere near as large as the gains for blacks; and consequently, the gap shrunk by a large amount.

Again, however, the picture was not completely positive. During the late 1970s, the 1980s, and into the 1990s, concentrations of urban poverty and the racial isolation of African American students increased especially in the North. Considerable research indicates that these factors have a negative effect on student achievement (Smith and O'Day 1991a).

School Characteristics

In our analysis, O'Day and I (1991a) considered a variety of school factors. Two factors, beyond the effects of desegregation in the South, deserve mention in this review. The first was an increased focus across the country on supplemental or compensatory education for children from low-income families between 1966 and the early 1980s. The policy emphasis started with the

Table 3. NAEP Long-term Trend Reading Assessments 1971–1988: North versus South, Public School Students Only

Northeast and Midwest States

	AGE 9						AGE 13						AGE 17					
Birth Cohort	**Test Year**	**White Scale Score**	**Black Scale Score**	**White Black Diff**	**Stand Gap**	**Stand SE**	**Test Year**	**White Scale Score**	**Black Scale Score**	**White Black Diff**	**Stand Gap**	**Stand SE**	**Test Year**	**White Scale Score**	**Black Scale Score**	**White Black Diff**	**Stand Gap**	**Stand SE**
1954													**1971**	295.3	248.4	46.9	1.02	0.19
1958							**1971**	265.9	229.8	36.2	1.01	0.07	**1975**	296.0	246.2	49.8	1.13	0.14
1963	**1971**	219.5	179.3	40.2	0.96	0.12	**1975**	265.0	230.1	34.9	0.98	0.06	**1980**	290.9	247.4	43.5	1.04	0.21
1967	**1975**	220.2	186.4	33.9	0.88	0.05	**1980**	265.5	233.0	32.5	0.93	0.06	**1984**	292.7	264.6	28.1	0.70	0.06
1971	**1980**	222.2	192.4	29.8	0.79	0.08	**1984**	260.5	238.2	22.3	0.63	0.07	**1988**	295.7	273.8	22.0	0.59	0.09
1975	**1984**	218.6	189.5	29.1	0.71	0.07	**1988**	260.7	240.4	20.3	0.58	0.10						
1979	**1988**	221.1	187.7	33.5	0.81	0.08												
1971–1988		**1.6**	**8.4**		**-0.14**			**-5.3**	**10.6**		**-0.43**			**0.4**		**25.4**		**-0.43**

Note: Northeast and Midwest states include CT, MA, NJ, PA, RI, VT, NH, ME, IL, KA, MI, MN, NE, OH, WI, IA, ND, and SD. South and Border states include AL, AR, FL, GA, LA, MS, NC, SC, TN, TX, VA, DE, DC, KY, MD, OK, WV, and MO.

NOTE: This table does not appear in Smith and O’Day (1991a).

SOURCE: Special tabulations, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Table 3. NAEP Long-term Trend Reading Assessments 1971–1988: North versus South (continued)

South and Border States

	AGE 9						AGE 13						AGE 17					
Birth Cohort	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE	Test Year	White Scale Score	Black Scale Score	White Black Diff	Stand Gap	Stand SE
1954													**1971**	290.4	237.2	53.2	1.16	0.08
1958							**1971**	256.4	218.8	37.7	1.05	0.06	**1975**	289.6	238.9	50.7	1.15	0.12
1963	**1971**	207.0	164.2	42.9	1.02	0.08	**1975**	260.0	224.8	35.3	0.98	0.06	**1980**	291.9	239.8	52.1	1.25	0.26
1967	**1975**	211.7	178.2	33.4	0.87	0.06	**1980**	260.9	229.3	31.6	0.91	0.13	**1984**	293.3	261.9	31.4	0.78	0.05
1971	**1980**	218.1	186.3	31.7	0.84	0.10	**1984**	262.0	232.5	29.6	0.83	0.07	**1988**	292.4	271.4	21.0	0.57	0.09
1975	**1984**	213.0	183.7	29.3	0.71	0.07	**1988**	259.7	243.3	16.4	0.47	0.13						
1979	**1988**	217.3	188.9	28.4	0.69	0.09												
1971–1988		**10.3**	**24.7**		**-0.33**			**3.2**	**24.5**		**-0.58**			**2.0**	**34.2**		**-0.60**	

Note: Northeast and Midwest states include CT, MA, NJ, PA, RI, VT, NH, ME, IL, KA, MI, MN, NE, OH, WI, IA, ND, and SD. South and Border states include AL, AR, FL, GA, LA, MS, NC, SC, TN, TX, VA, DE, DC, KY, MD, OK, WV, and MO.

NOTE: This table does not appear in Smith and O'Day (1991a).

SOURCE: Special tabulations, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1971 to 1996 Long-term Trend Reading Assessments.

Great Society and eventually characterized much of the federal education effort. Title I of the Elementary and Secondary Education Act was the cornerstone program of this effort, providing supplemental assistance to millions of students. Title I programs provided substantial resources and focused attention on the need, particularly in high-poverty schools, to provide low-achieving students with the kinds of support that they needed to learn the "basics."

It is difficult to disentangle the effects of Title I from the effects of other programs, primarily because the program served the universe of high-poverty elementary schools. There are no data sets that show large, specific effects for the Title I program during this period; nevertheless, the program put substantial supplemental educational resources into schools, helped in many districts to lower class sizes, and emphasized the basics of reading and mathematics (Kaestle and Smith, 1982). Perhaps more important, Title I served as a national stimulus and symbol to focus attention on the needs of low-income students. Many Title I students were African American. In my view, it was no coincidence that the growth and expansion of Title I and of "the national idea" of supplemental education support came at the same time as substantial increases in African American scores on NAEP reading and mathematics assessments at the 9- and 13-year-old levels.

These changes in curriculum and in schooling were largely confined to the elementary and middle school levels; however, substantial decreases in the achievement gap between black and white students also occurred at the secondary level from the mid-1970s through the late 1980s. O'Day and I (1991a) argued that this happened, in part, because many secondary schools were also focused on compensatory education. The stimulus in this case came from minimum competency exams. By the mid-1980s, 33 states required passage of a minimum competency test of basic skills as a criterion for graduation (Office of Technology Assessment 1992). We believed that the use of minimum competency assessments had a powerful ameliorative influence on the achievement gap. The instructional emphasis on basic skills, combined with high-stakes testing, produced a greater degree of focus and coherence in the core curriculum of many secondary schools, which might have been lacking in prior years.

Note that, at both elementary and secondary levels, the school-based policies that appear to have affected the gap were focused on basic rather than advanced skills. This focus is consistent with the effects on achievement results during this time period. In summary, the effects were generally positive for most students who started the era with relatively low scores and neutral for students who started with higher scores. Further, these results are also consistent with the character of the NAEP trend assessment instrument, which itself measures basic skills primarily rather than more complex intellectual abilities.

The Gap Ceases to Narrow: Beyond 1988

Through the 1988 test administration, the NAEP trend data for reading and mathematics achievement showed a powerful and consistent reduction in the gap between the scores of African American and white students at all three age levels. From observation of these achievement data, there was no reason to believe that this trend would not continue. Yet O'Day and I (1991a) argued that the gap would stop narrowing unless some new policies were quickly put into place.

Our argument then was based on three points. First, each of the variables that we believed contributed to the reduction of the gap in black and white achievement had changed by 1990. For example, poverty was slightly increasing during the late 1980s for black families, rather than decreasing. Rising poverty rates would tend to reduce rather than increase average black achievement. Another example is that desegregation and the migration of black families North and to the cities was over; and the concentration of poverty in the Northern inner cities probably was making a negative, rather than positive, contribution to reducing the gap. Finally, the number of states using minimum competency tests as a graduation requirement had stopped increasing.

Our second point was that during the 1980s the national idea of assisting the poor through government programs was under attack. The focus on *equality* had diminished and had given way to a new focus on *quality*. The 1983 report, *A Nation at Risk*, had revealed the inadequate performance of American students compared to their international peers (National Commission on Excellence in Education 1983). This report stimulated the already growing interest among the states, especially in the South, in reforming the schools through such measures as increasing the number of hours in a school day, adding to the number of days in a school year, and encouraging students to take tougher courses.

While an overall increase in quality would support children from low-income as well as affluent families, two observations indicated that these reforms would not reduce the gap. One observation is that for years the United States had condoned a curriculum and a quality of instruction in low-income inner-city and rural schools that are shallow and insufficient compared to those available to their suburban and well-to-do peers. There was little in the new *quality* reforms that would redress this *inequality* in opportunity. Another observation is that O'Day and I realized that, even if there were significant improvements in curriculum and instruction, they would take considerable time to implement and, further, that schools in inner cities and poor rural areas lag their counterparts in the suburbs in carrying out such changes. Thus, even though we hoped we were incorrect, we suggested that poor and minority students were less likely to benefit from the new focus on higher quality, at least in the short run.

Finally, the third point was that, while black students had shown striking increases in their scores between 1971 and 1988, there was reason to believe that many of the gains on the assessments that could be achieved from across-the-board emphases on the basics and minimum competencies had already been achieved by the late 1980s. Thus, in the future, closing the gap would require greater attention to higher-order skills such as reading comprehension and problem solving than poorly performing students had received in the past.

On the basis of these conjectures, O'Day and I questioned whether the gap would continue to reduce in size over the decade of the 1990s (Smith and O'Day 1991a). We suggested that it would *not* continue to narrow unless three vigorous steps were taken to differentially improve the quality of educational opportunity for African Americans. In particular, we argued that, in the absence of an effective policy to alleviate poverty, the country needed an aggressive effort to support low-income families with children and to prepare all children for school by improving their access to quality health services and early childhood education.

Outside of schools, we advocated increasing the opportunities for school- and community-based after-school programs for students in inner-city and poor rural areas, in order to provide safe and academically stimulating environments beyond the six-hour school days. Within schools, we argued for rapid movement toward state standards-based reform, a relatively new idea then, though well understood now.[7] Our argument then centered on the need to eliminate the

[7] See Smith and O'Day (1991b) for additional details.

decades-long practice of giving our neediest students in the highest poverty schools the least-trained teachers and a “watered-down” curriculum. The curriculum and the quality of instruction needed to be upgraded throughout many schools, but the need was greatest in high poverty areas. I will return to these policies in my discussion of how to address the achievement gap, in the last section of the paper.

Now, what happened after 1988? In 1992, O’Day and I wrote a second paper in which we specifically examined the results of the 1990 assessment (O’Day and Smith 1993). To understand what happened to the achievement gap in 1990, we need to refer once again to table 1 (p. 252). The most dramatic changes happened in reading achievement. For the 17-year-olds in the 1971 cohort, who were tested in 1988, the test score difference between African American and white students was 20 points, down 32 points compared to 1971. Then, suddenly, in 1990 the gap enlarged to 29 points. For 9-year-olds the gap increased by 6 points, and for 13-year-olds by only 2 points. For mathematics achievement at the 9- and 13-year-old levels there was little change in the gap from 1986 to 1990, while for the 17-year-olds the gap narrowed somewhat. Now, of course, this was only one new point in time, and one data point is clearly insufficient for making strong inferences about changes in a long-term trend. Still, for reading achievement, the 1990 assessment revealed sudden, substantial changes in the size of the achievement gap in precisely the direction that we feared. Further, for two of the age levels in mathematics, the gap remained fairly constant between 1986 and 1990 (see table 2, p. 253).

We now have the luxury of looking over a longer period of time. As a number of authors in Jencks and Phillips (1998) suggest, the NAEP longitudinal trend data from 1990 through 1996 do not show a clear pattern of growth for either African American or white students in either reading or mathematics achievement. Consequently, for these trend data there is no clear pattern of change in the gap during the early and middle 1990s. It seems clear that the gap had ceased to reduce, at least as measured by the NAEP trend data. But the NAEP trend data may not be telling the entire story.

The 1990s: Standards-Based Reform

During the decade of the 1990s, the nation’s focus on educational improvement and reform was unprecedented. Federal, state, and local governments made education their top priority. At the federal level, three fundamental objectives guided investment strategies in K–12 education, as follows:

1. Create economic and health care environments as stable and livable as possible for all families with children.
2. Expand opportunities for all students to participate in engaging and educationally rich activities beyond the traditional school hours. These opportunities include high-quality preschool and after-school opportunities, particularly for children from the least affluent families.
3. Stimulate and support state and local standards-based reform strategies to improve the quality of schools for all students.

For each of the objectives, substantial progress has been made by the Administration, Congress, and the states, though there continues to be considerable distance between current conditions and the fulfillment of these ambitious goals. For the first objective, sustained economic growth throughout the middle and late 1990s created over 10 million new jobs, and the unemployment rate has dropped to record lows. In addition, the Earned Income Tax Credit (EITC) provided millions of low-income families with children additional income to lift them above the official poverty level; and the Children's Health Insurance Plan (CHIP) makes it possible for every child in a low-income family to have adequate health care.

The second objective has seen the development of education standards for the Head Start curriculum and the expansion of Head Start enrollment, which has brought 72,000 children into the program since the reauthorization in 1994.[8] In addition, the Administration has dramatically expanded the 21st Century After-School Program, from a $1 million program in FY1996 to a $450 million program in FY2000 that will serve well over half a million students during the school year 1999-2000. Both of these programs are targeted to provide services to low-income students who would not otherwise have either preschool or after-school educational opportunities.

Finally, the nation's standards-based reform movement is based on the principle of establishing coherent and fair policies at the state and local levels to improve student learning (Smith and O'Day 1991b). Standards-based re-

[8] See www2.acf.dhhs.gov/programs/hsb.

form has taken on a life of its own and is now the dominant reform in most states—it has become a national idea. There are four fundamental elements of this reform:

1. Establish challenging state content and performance standards for all students.
2. Align all parts of the education system to assist all students to learn the content, skills, and strategies set out in the state standards. The curriculum, teacher training, technical assistance, and assessments should all be based on the state content and performance standards.
3. Provide local school districts and schools with sufficient fiscal resources and at the same time grant sufficient flexibility, autonomy, and responsibility to the districts to use their resources to maximize students' opportunities to achieve to the standards.
4. Develop and implement accountability systems that use student performance measures to demonstrate to the public that schools and districts are meeting their obligations to teach all students to challenging content standards.

This reform strategy has shaped federal and state education policies throughout the decade. At the federal level, Goals 2000, a separate program proposed by the Clinton Administration and passed in 1994, provides support to states to develop and implement standards-based reforms. The core federal programs in education (including Title I, the program for supplemental services for low-achieving students) were reauthorized and modified to support the state reforms. Beginning in 1994, Title I legislation required that all students eligible for Title I services be taught to the same challenging state standards as all other children in the state—this requirement focuses specifically on closing the gap. This standards-based reform strategy has provided a focus for a variety of other supportive interventions, including the use of technology in the schools, the professional development of teachers, and even the stimulation of charter schools.[9]

In 1993, few states had any type of coherent content standards. In 1998, 44 states had content standards in at least three subjects (*Education Week* 1998).

[9] Secretary of Education Richard Riley, in his State of American Education Address (February 22, 2000), stated the number of charter schools as about 1,700. See www.ed.gov.

Most states have their assessments aligned with these content standards (*Education Week* 1998). In most districts, the curriculum and the textbooks are examined and selected based upon their alignment to the state standards. A substantial number of states are also establishing programs of teacher training and teacher professional development designed to prepare teachers to teach to these new content and performance standards. Though it may take several years to achieve full implementation in many states, I believe that the cumulative force of such coherent strategies and policies in so many states is already having positive effects on student achievement.

New Assessment Data from NAEP: Three Policy Perspectives

To help explore this belief, I want to introduce a new set of NAEP data into the conversation—data from the "main" NAEP. The NAEP trend data presented in the previous sections of this paper came from a supplemental assessment that has been administered since 1971 strictly for the purpose of maintaining longitudinal trends. This NAEP trend assessment has undergone few changes in format or content since 1971. As a consequence it does not measure some of the concepts, knowledge, and skills that have been introduced into the curriculum in more recent years.

Another NAEP assessment—called the main NAEP—serves now as the primary measure of the achievement of the nation's students. Since 1990, the main NAEP has measured a wider range of skills and knowledge with a broader repertoire of item types than the trend NAEP. The 1990 version is administered on a regular schedule at the national level and also on a voluntary basis by many states. The main NAEP data show promising signs of the effectiveness of the state standards-based reform movement. Three views of data from the main NAEP are illustrated in tables 4, 5, and 6 below.

Reading Achievement

Table 4 sets out national results for reading at three grade levels (fourth, eighth, and twelfth) for whites, African Americans, and Hispanic Americans. Reading scores were collected in each of three years: 1992, 1994, and 1998. During the period 1992–1994 the effects are generally negative; with one exception the scores dropped over those years, and the losses were greater for African Americans and Hispanic Americans than for whites.

However, from 1994 to 1998 the reading scores for all groups increased; and the increases were slightly larger for African Americans and Hispanic Americans than for whites. Why is there a change in direction of these scores? One hypothesis is that the 1994 test administration is an anomaly—that it somehow represents a mismeasurement—and that we should therefore overlook the results and simply note the difference between the 1992 and the 1998 results. For grades fourth and eighth, this comparison reveals some slight increases over the 6 years, except for fourth grade Hispanic Americans; and for grade 12, slight decreases.

There is a second possibility. Since Goals 2000 did not pass until 1994 and many states did not begin their reforms until then, one could argue that 1994 is a better baseline year than 1992 for estimating the effects of the new state reforms. The gains from 1994 to 1998 are large for all groups at all grades, suggesting that the reforms may be having a positive effect. Whatever the case, the scores are now moving in the right direction. Note, however, that there is no evidence of the black-white gap's closing.

Mathematics Achievement

The overall picture is more optimistic for mathematics, as seen in table 5. These data show strong gains in achievement for all but one of the comparisons from 1990 to 1992 and from 1992 to 1996. Unfortunately, we do not have 1994 or 1998 data for mathematics achievement. Thus there is no true baseline data point to measure the effects of the reforms. Overall, from 1990 to 1996, the gains are more than 1.0 grade levels for all but three of the nine comparisons in table 5; for those three, they are between 0.5 and 1.0 grade levels. The effects here are difficult to attribute to standards-based reform because of the lack of a good baseline, but they are consistent with the possibility of a positive effect from the reforms.

Thus, the picture for the main NAEP data for the 1990s indicates a potentially positive effect of the state standards-based reforms on both reading and mathematics achievement. These data by no means present an ironclad argument, but they are suggestive. Again, there is no sign of a reduction in the gap.

State Trends

A second perspective views the results from the state NAEP as additional support for the argument concerning the effectiveness of the reforms. David

Table 4. Reading Achievement: NAEP National Data by Year, Grade, and Racial-Ethnic Group

	1992	1994	1998	Difference in Scale Scores	
				1994–1992	1998–1994
4th Grade					
White	225	224	227	-1	+3
African American	193	187	194	-6	+7
Hispanic	201	191	196	-10	+5
8th Grade					
White	267	268	272	+1	+4
African American	238	237	243	-1	+6
Hispanic	241	240	244	-1	+4
12th Grade					
White	298	294	298	-4	+4
African American	273	265	270	-8	+5
Hispanic	278	270	275	-8	+5

NOTE: Includes both private and public school students.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992, 1994, and 1998 Reading Assessments.

Table 5. Mathematics Achievement: NAEP National Data by Year, Grade, and Racial-Ethnic Group

	1990	1992	1996	Difference in Scale Scores	
				1992–1990	1996–1992
4th Grade					
White	220	228	232	+8	+4
African American	189	193	200	+4	+7
Hispanic	198	202	206	+4	+4
8th Grade					
White	270	278	282	+8	+4
African American	238	238	243	0	+5
Hispanic	244	247	251	+3	+4
12th Grade					
White	301	306	311	+5	+5
African American	268	276	280	+8	+4
Hispanic	276	284	287	+8	+3

NOTE: Includes both private and public school students.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1990, 1992, and 1996 Math Assessments.

Grissmer of RAND was asked by the National Goals Panel to look behind the results of state NAEP and to try to explain why North Carolina and Texas performed better compared to other states during the 1990s (Grissmer and Flanagan 1998). Grissmer and his colleague, Ann Flanagan, found, among other things, that both North Carolina and Texas have implemented standards-based reforms in several subject areas over the last few years. These states maintain consistent policies that emphasize relatively challenging standards, require curriculum-aligned tests, provide for accountability at the school level, offer extensive teacher training, and focus special efforts on low-scoring students. In addition to the focus on teaching and learning, Grissmer and Flanagan found that the financial support and committed involvement of the business community and the sustained focus on education in government, despite partisan shifts in the political leadership, were also positive influences on the effectiveness of the reforms.

To illustrate the effects of such coherent policies, table 6 presents state NAEP data on fourth graders from the 1992 and 1996 mathematics assessments at two achievement performance levels, making possible comparisons between state-level and nationwide data. Nationwide, the percentage of white students who scored at or above the Basic (the first level) performance level increased from 69 percent in 1992 to 74 percent in 1996. For blacks nationwide, the percentage of students scoring at or above Basic increased from 22 percent in 1992 to 32 percent in 1996. For Hispanic American students nationwide, the increase in students scoring at or above Basic was from 33 percent in 1992 to 40 percent in 1996. In other words, each group improved; and, on this measure, minority fourth graders across the country may be once again beginning to close the achievement gap, though there is still a long way to go.

Table 6 shows clearly the difference between the results in Texas and North Carolina compared to the national results and also to student performance in three other states that serve as rough benchmarks—California, Florida, and New York. In Texas, which has focused intense efforts on improving performance in low-scoring schools, white fourth graders scoring at or above Basic moved up from 72 percent in 1992 to 85 percent in 1996. For blacks, the increase at or above Basic was from 29 percent to 47 percent. For Hispanic American students, the percentage increase was from 43 percent to 55 percent. The data for North Carolina reveal similar increases for both black and white students. For whites, the percentage achieving at or above the Basic level went

Table 6. 1992 and 1996 Mathematics National Assessment, Percentage of Students at *At Or Above Basic* Achievement Level by Race-Ethnicity, Grade 4 Public School Students

	White		Black		Hispanic	
	1992	**1996**	**1992**	**1996**	**1992**	**1996**
Texas	72	85	29	47	43	55
North Carolina	65	77	24	37	35	43
California	61	63	21	18	27	29
Florida	66	70	22	26	27	29
New York	71	80	31	37	33	40
Nation	**69**	**74**	**22**	**32**	**33**	**40**

SOURCE: These scores are from the main NAEP assessments, not from long-term NAEP trend assessments.

from 65 in 1992 to 77 in 1996; for blacks, the percent scoring at or above Basic increased from 24 in 1992 to 37 in 1996.

In contrast to North Carolina and Texas, California had a comparatively incoherent and weak strategy for school reform during the early and middle years of the 1990s, and the numbers in table 6 seem to track that incoherence. For white students, the percentage achieving at or above the Basic level in California barely moved upward from 61 in 1992 to 63 in 1996. For black students, the percentage dropped from 21 to 18, and for Hispanic students, the percentage moved up a small amount, from 27 to 29. A comparison of the data between California and Texas (which is somewhat similar to California in the size and diversity of its student population) underscores the fact that, for one reason or another, some states did well in their efforts to improve student outcomes, while others did not.

State Reform Rankings and State NAEP Data

A third and admittedly speculative perspective on examining the effects of the reforms uses the state NAEP data for reading for 1994 and 1998. The trade newspaper, *Education Week*, published an analysis of the progress of the various states on standards-based reform as a stand-alone insert to its publication (*Education Week* 1998). As part of its analysis, *Education Week* ranked the quality of the state reforms according to a number of criteria that represented whether they had achieved certain components of the reform up to that time.

I hypothesized that, if the rankings were valid and the reforms were having an effect on achievement, then the ranking of the reforms should correlate with the gains that the states made on NAEP. I used fourth grade reading gains by state over the period 1994–1998. Because only 33 states had administered the state NAEP in both 1994 and 1998, the sample was somewhat constrained. Nonetheless, after controlling for differences in state per pupil expenditures, the partial correlation between the rankings and the gain scores was +0.43; after controlling for state poverty levels, the partial correlation was +0.46. In each case the partial correlation was statistically significant at the .05 level on a one-tailed test of significance.

Final Remarks: Research Priorities and Methodological Issues

A major purpose of this conference is to examine methodological issues and their inter-relationship with the capability of education research to understand the perplexing questions regarding the test score gap. The importance of these issues is intensified because the dynamics of schooling, and possibly the causes of the achievement gap, are changing. The extraordinary developments in information technology, a new and demanding market economy, the concentration of sustained poverty among a large percentage of families with children (particularly African American and Hispanic American families), and a wide variety of immigrants speaking many different languages are transforming both our expectations and our requirements for schooling.

In this context, I would like to explore themes in two areas. The first theme is the use of experiments to provide more reliable and valid research information about how to help close the achievement gap. The second area has to do with measurement concerns that emerge in thinking about the achievement gap, such as how we measure performance against a standard. Making progress in each of these areas is important to creating a strong foundation of evidence upon which solid policy can be built.

The Use of Experiments

Two general points are important when evaluating research. First, a good model grounded in theory should facilitate explanations of results, and carefully designed measures should provide clear understanding of the data in any study. The second point is that every methodology has weaknesses and strengths and that, further, it is important to understand these weaknesses and strengths in order to select either a single methodology or a combination of different methodologies to examine the research question. The use of true experiments should be a significant part of the repertoire of federal evaluation programs. Random assignment and deliberate intervention, taken together, provide a useful tool to test hypotheses and estimate effects, whether a theory is strong or weak. When they are appropriate, and if carefully designed and implemented, controlled field experiments can often demonstrate powerful and persuasive evidence.

The potential strength and authority of experimental field trials has been demonstrated by two sets of studies. The first is the set of Tennessee STAR (Student Teacher Achievement Ratio) studies (Finn and Achilles 1999; Word, Johnston, and Bain 1990). The Tennessee class size reduction study in the early grades, initiated over a decade ago, yielded findings that have had an enormous impact on policy debates across the nation. In addition to recent federal legislation (that funds the hiring of additional teachers), several states and local districts are moving toward smaller classes, particularly in the early grades. Nearly all of these legislative initiatives are motivated, at least in part, by the findings of the Tennessee experiments. The fact that this study included randomization of students to classes of different sizes contributed greatly to its influence on policy discussions. Its acceptance was also enhanced by the large number of schools involved in the experiment and by the commonsense nature of the treatment (small class size in the early grades to improve student achievement). Further, the reputation and importance of the STAR study were enhanced by follow-up studies that showed that the initial positive results were sustained in the middle and high school years (Nye et al. 1994).

Another example of the use of experimental methodology can be found in the series of studies on early reading acquisition conducted by the National Institute of Child Health and Human Development (NICHD) under the direction of Reid Lyon (Moats and Lyon 1997; Snow, Burns, and Griffin 1998). These studies consisted of interlocking experiments, conducted over a 10-year period, that examined a set of theoretical hypotheses related to word recognition and reading in young children. The experiments were developed in such a manner that the findings from one study were linked to related hypotheses in other studies. Just as with the STAR study, the fact that the studies were experiments added to their credibility with Congress. The reputation of the NICHD as a more “scientific” agency, rightly or wrongly, was another factor in the credibility accorded this work. Credibility is often a necessary first step, but it is not enough to ensure that the findings are utilized. The NICHD researchers buttressed their credibility with a commonsense approach and detailed descriptions of their findings in a coherent and compelling fashion. Their story is about how their related set of experiments provided quality insights that can improve the chances of many young children to learn to read. These two examples demonstrate both the feasibility of field experiments and their potential influence in the policymaking process.

As we consider a research agenda for OERI and NCES in the future, several straightforward recommendations from this research seminar for experimental studies could be productive in the clarification of programmatic interventions that can narrow the achievement gap. For example, in the area of promoting better quality in the preparation of students for school, one possibility would be to conduct experiments that address different approaches for training parents to use new understandings of the development of cognitive processes to help their own children. Within schools, the NICHD studies could be extended to cover large and more diverse settings to determine the strength of the reading interventions in less controlled environments. Along the same line, a thoughtful set of experiments that explore the implications of the NICHD research for Limited English Proficient (LEP) children could be important. In an emerging area, literally dozens of potentially powerful field experiments need to be conducted on some of the promising new ways of using technology in classrooms and for distance learning. In this arena, we will need to learn how to carry out experiments in the “real time” necessitated by the rapid changes in the nature of technology. Finally we need experimental data on the effects on students of summer school and after-school programs. Early experiments in the 1970s found few effects for summer schools—these questions need to be revisited as summer schools become more and more a method of expanding students’ opportunities to learn.

Methodological Issues in the Measurement of the Gap

Finally, let me suggest two important methodological issues in the measurement of the gap in performance between white and minority students. First, OERI and NCES need to develop a research agenda aimed at understanding what the two types of NAEP assessments (trend and main) *actually* measure. In other words, current efforts to assess student achievement would benefit from carefully conducted construct validity studies. The data from the two assessments reveal quite different results during the 1990s—it would be nice to know why.

Second, at the present time, various states are struggling toward a set of more sophisticated assessments. These assessments should be aligned with state content standards that set out the skills and knowledge that students should know and be able to do. Also, the assessments should be designed with performance standards or levels that indicate how well a student has learned the skills and knowledge. For example, where a content standard might specify that a student should be able to write a short persuasive essay, performance standards would provide a way of measuring the quality of the persuasive essay. If done well, performance standards would not be established by selecting cut scores on norm-referenced assessments as the NAEP achievement levels are now defined. As assessments improve, performance standards should become real in some sense. For example, reading at or above the Basic level could be validated to show that a student can read and comprehend a well-defined set of books and passages. Similarly, more demanding performance items could define and assess other more challenging levels beyond Basic.

Well-constructed content and performance standards would be the products of the intersection of reasonable theories of the content area, human development, human learning, and pedagogy. Growth through performance levels might be discontinuous rather than smooth. For example, achieving a Basic level of reading in the fourth grade may require effective word attack skills, while achieving a higher level may require mastery of strategies of comprehension. That is, a different dimension of mastery may be required to achieve a higher level, one involving a qualitatively different set of skills and knowledge. This would have substantial implications for the curriculum, as well as for the interpretation of group differences. In these circumstances, the achievement gap would no longer be measured by a scale score difference, but by differences in the percentages of students achieving to the different performance levels. Measurement of the gap becomes more complicated conceptually and methodologically because there are multiple comparisons related to the different performance levels. The problems are already apparent when the trends from the new NAEP assessments and the results of some state assessments are interpreted.

Until we have content and performance standards established on the basis of reasonable theory, we will continue to have political and educational problems with how to set them. A current political issue has to do with how challenging is "challenging," or how do we determine where to set performance levels so that they positively rather than negatively motivate students and educators to succeed? A goal considered impossible to attain by students, teachers, and parents may undermine the credibility of the reforms. Moreover, the size of the achievement gap, as measured using a performance level, may be determined in part by how high the bar is set, that is, by the degree of difficulty of the performance standard. A bar or performance standard that is set relatively low will have higher percentages of students passing and, therefore, will typically result in a smaller gap between groups that have different levels of command of the tested content and skills. Conversely, a bar set very high may make it practically impossible to have high percentages of students pass and may exaggerate the magnitude of differences among groups. This is not an academic issue. In the past the tendency has been to lower the bar to minimize failure. This approach, however, fails to meet the purpose of the reforms to challenge all students to meet rigorous standards. But, now, there also are a number of key states, including New York and Massachusetts, where thoughtful analysts believe that the bar has initially been set too high, and this policy may jeopardize support for the reforms.

Conclusion

Educational excellence and educational equality—again I emphasize that these are the nation's co-equal guiding goals. These are the overarching ideals that researchers and policymakers must continuously keep in mind in their consideration of such important issues as which interventions most effectively enhance student learning, which state systemic structures support achievement in a cost-effective manner, and which collaborations and partnerships can achieve strong public and parental support, while advancing overall improvement in schools in the long term. My hope is that, in the deliberations of this seminar, we will find that we have begun to chart a course of recommended research directions and policy alternatives that will, both sooner and later, assist us in solving the remaining challenges, increase our commitment to the goals, and eventually help in the attainment of those national ideals.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Council of Chief State School Officers. (1994, May). *Baselines for Goals 2000 Implementation.* Washington, DC: Author.

Education Week 18(17). (1998, January 11). Quality Counts. Bethesda, MD: Author.

Finn, J. D., and Achilles, C. M. (1999, summer). Tennessee's Class Size Study: Findings, Implications and Misconceptions. *Educational Evaluation and Policy Analysis* 21 (2).

Grissmer, D. W., Kirby, S. N., Berends, M., and Williamson, S. (1994). *Student Achievement and the Changing American Family*. Santa Monica, CA: RAND.

Grissmer, D. W., Williamson, S., Kirby, S. N., and Berends, M. (1998). Exploring the Rapid Rise in Black Achievement Scores in the United States, 1970–1990. In U. Neisser (Ed.), *The Rising Curve: Long-term Changes in IQ and Related Measures.* Washington, DC: American Psychological Association.

Grissmer, D. W., Flanagan, A., and Williamson, S. (1998). Why Did Black Test Scores Rise Rapidly in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.

Grissmer, D., and Flanagan, A. (1998, November). *Exploring Rapid Achievement Gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel. (www.negp.gov/Publications/LessonsfromtheStates).

Jencks, C., and Phillips, M. (1998). (Eds.) *The Black-White Test Score Gap*. Washington, DC: The Brookings Institution Press.

Kaestle, C. F., and Smith, M. S. (1982, November). The Federal Role in Elementary and Secondary Education, 1940–1980. *Harvard Educational Review 52*(4): 384–408.

Moats, L. C., and Lyon, R. (1997, November–December). Critical Conceptual and Methodological Considerations in Reading Intervention Research. *Journal of Learning Disabilities 30*(6): 578–588.

National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperatives for Educational Reform*. Washington, DC. Author.

Nye, B., Zaharias, J., Fulton, B., Achilles, C. M., and Cain, D. (1994). *The Lasting Benefits Study: A Continuing Analysis of the Effect of Small Class Size in Kindergarten through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels.* Unpublished paper. Tennessee State University, Nashville.

Office of Technology Assessment. (1992). *Testing in American Schools: Asking the Right Questions*, p. 15. Washington, DC: U.S. Government Printing Office.

Schofield, J. W. (1991). School Desegregation and Intergroup Relations: A Review of the Literature. In G. Grant (Ed.), *Review of Research in Education*, *Vol. 17* (pp. 335–409). Washington, DC: American Educational Research Association.

Snow, C. E., Burns, M. S., and Griffin, P. (Eds.) (1998). *Preventing Reading Difficulties in Young Children*. Committee on the Prevention of Reading Difficulties in Young Children, National Research Council. Washington, DC: National Academy Press.

Smith, M. S., and O'Day, J. A. (1991a). Educational Equality: 1966 and Now. In D. Verstegen and J. Ward (Eds.), *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook.* New York: Harper Business.

Smith, M. S., and O'Day, J. A. (1991b). Systemic School Reform. In S. Fuhrman and B. Malem (Eds.), *The Politics of Curriculum and Testing: The 1990 Yearbook of the Politics of Education Association* (pp. 223–276). London: Falmer Press.

Smith, M. S., and O'Day, J. A. (1993). Systemic Reform and Educational Opportunity. In S. Fuhrman (Ed.), *Designing Coherent Education Policy* (pp. 250–312). San Francisco, CA: Jossey-Bass Publishers.

Word, E., Johnston, J., and Bain, H. (1990). *Student Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study: Final Summary Report 1985–1990.* Nashville: Tennessee Department of Education.

Educational Research and Educational Policy: An Historical Perspective

Christopher Jencks
Malcolm Wiener Center for Social Policy
John F. Kennedy School of Government
Harvard University

When I began doing quantitative research on education in the mid-1960s, most of what social scientists thought they knew about the effects of educational policies was based on experiments. There were not many experiments, and they did not all reach consistent conclusions, but they were about all we had to go on.[1] School administrators and teachers paid this research very little heed. When they thought about the effects of different educational policies, they based their judgments on personal experience. Educators who had spent years in the schools mostly had strong views about what worked and what did not. Of course, educators often disagreed with one another about what lessons experience taught, but these disagreements seldom led to self-doubt. Nor did educators who disagreed seek to resolve their disagreements by reading educational research.

Although educators seldom sought researchers' advice, researchers continued to offer it. They ran small experiments that tried to assess the accuracy of educators' beliefs about such matters as class size, ability grouping, and the best way to teach reading. More often than not, researchers interpreted their findings as showing that educators' beliefs were wrong. As a result, educational researchers tended to feel superior to those who staffed the schools.

The 1960s: Misinterpreting Insignificant Coefficients

Although educational researchers often felt superior, educators seldom felt inferior. Faced with a study showing, let us say, that children's spelling

[1] The opinions expressed in this paper are those of the author. Address any comments to Professor Christopher Jencks, John F. Kennedy School of Government, Malcolm Wiener Center for Social Policy, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138.

skills did not improve over the long run if teachers handed out a list of words every Monday and gave a spelling test every Friday, most educators simply dismissed the findings as implausible. Likewise, when studies seemed to show that children learned no more in small classes than in large classes, few educators considered the possibility that the studies might be right. In most cases educators did not even bother to dismiss such results as implausible. They simply ignored educational research entirely. Legislators did the same thing.

Educators' indifference to research results convinced many researchers that the people who staffed the nation's schools were unscientific traditionalists, unwilling to consider the possibility that their prejudices were ill founded. In retrospect, however, the educators' indifference to educational research seems largely justified.

Then as now, educational experiments typically assessed policies by comparing outcomes for students who had had different educational experiences. Sometimes these comparisons involved small experiments in which students were randomly assigned to different treatments. Sometimes they involved "natural experiments," in which students had different educational experiences because school boards, principals, or teachers followed different practices. Since there were no national or statewide testing programs that allowed researchers to link students' achievement to past experiences, most of these comparisons were based on small samples. As a result, the sampling errors of the estimates were usually quite large, and the difference between those who had had different experiences was often less than twice its sampling error. Researchers almost always interpreted this finding as supporting the "null hypothesis," namely that the experience in question made no difference.

With the wisdom of hindsight, this interpretation of insignificant coefficients looks foolish (though it is still disturbingly common). Every first-year statistics student learns that data analysts can make two different sorts of errors. "Type One" errors occur when the analyst accepts a false hypothesis as true. "Type Two" errors occur when the analyst rejects a true hypothesis as false. Social scientists have traditionally been far more concerned about avoiding Type One than Type Two errors. This bias makes sense when social scientists are testing their own theories. It may also make sense when social scientists are testing theories of interest only to other social scientists, since most such theories are too simple to be useful and reducing intellectual clutter is always a high priority in science.

In the policy arena, however, most data analysts are testing hypotheses that are widely accepted in the real world, not hypotheses that come from some theorist's fevered imagination. This reality means that when samples are small and measurement imprecise, policy researchers who emphasize significance tests are far more likely to reject a true hypothesis than to accept a false hypothesis. If practitioners were to take such researchers' conclusions seriously, they would often be led badly astray.

With the wisdom born of hindsight, one can see that policy researchers had several better alternatives. Bayesian theory suggests, for example, that researchers should start out by formulating "priors" that describe their best guesses about how the world works. As they accumulate additional evidence, either from experiments or other sources, they should update their priors to incorporate this new information. Had educational researchers tried to proceed in this fashion, their priors about policy questions would presumably have been shaped by two considerations, as follows:

1. If most educators think that a policy enhances student achievement, and if there is no other evidence about the policy's impact, a reasonable person should assume that practitioners are somewhat more likely to be right than wrong.
2. If an educational policy had *very large* effects, this would be obvious to everyone, and we would not be doing research on the policy's impact. Thus, if we are doing research on a policy's effect, the effect is not only likely to be positive, but also likely to be relatively modest.

If researchers had reasoned in this way, they would hardly ever have started out with the null hypothesis—the theory that a popular policy has no effect whatever. Thus, when they generated new data showing that a policy's impact was quite uncertain, they would not have raised the possibility that the policy had no effect. Instead, they would have concentrated on estimating the actual size of the effect.

For educational researchers who had no prior expectations about how large an impact a policy was likely to have, traditional statistical methods offered another attractive option. Researchers could just have reported the odds that a policy had a positive rather than a negative effect. Suppose, for example, that an investigator had randomly assigned first graders to one of two reading classes: a class of 15 and a class of 25. Suppose, too, that at the end of the year the children in the smaller class scored 0.3 standard deviations higher on the investigator's reading test than the children in the larger class, but the sampling error of this difference was also 0.3 standard deviations. The 95 percent confidence interval for the effect of being in a reading class of 15 rather than 25 therefore runs from –0.3 standard deviations to +0.9 standard deviations.

The most frequent interpretation of such a result is that since the confidence interval includes 0, we cannot reject the hypothesis that class size has no effect on reading achievement. A more plausible conclusion, I would argue, is that since five-sixths of values in the confidence interval are positive, the odds are 5 to 1 that small classes raise reading achievement rather than lowering it. If researchers have strong priors, of course, the odds that small classes raise reading achievement are even higher. The odds that class size has an impact of exactly 0 are, in contrast, vanishingly small.

Thirty years ago, however, researchers almost always emphasized statistical significance when formulating their conclusions. Even today, educational researchers who get statistically insignificant coefficients are more likely to suggest that the variable in question has no effect than to conclude that their sample was too small to justify any firm conclusions. Under these circumstances, I think that educators' skepticism about educational research was largely justified. Unfortunately, educators seldom had enough statistical expertise to explain their skepticism in a technically compelling way. So they just ignored educational research or dismissed it as irrelevant to classroom practice.

The Coleman Report

Educators' refusal to take educational research seriously faced its first important challenge in 1966, when the U.S. Office of Education released a report by James Coleman et al. analyzing the determinants of student achievement in some 4,000 schools throughout the country. Coleman and his colleagues found a weak relationship between many popular educational policies and student achievement, which was nothing new. But Coleman's report differed from earlier studies in several crucial respects. First, Coleman was a distinguished sociologist, whom the Office of Education had selected to carry out a Congressionally mandated study. Second, his analyses covered far more schools and students than any earlier study, so the results could not be dismissed as a fluke. Third, Coleman's work appeared at a time when the federal government was beginning to play an expanded role in educational agenda-setting and when policymakers were more attentive to the findings of social science than they had been in earlier periods. Fourth, when other social scientists reanalyzed Coleman's data, as many did over the ensuing years, they often faulted his methods, but usually came to broadly similar substantive conclusions.[2]

If educational researchers had been committed Bayesians, they might once again have said: "Well, the estimated effects of educators' preferred policies may be statistically insignificant, but the standard errors of the estimates are so large that we should not draw any strong policy conclusions from them." But we did not say that. I myself once tried to go down that road, and it was a dead end.

My first paper on the relationship between school policy and student achievement was a reanalysis of the Coleman data (Jencks 1971). In it, I reported the confidence interval for each estimated effect. As far as I know, nobody read this paper except the editors of the volume in which it appeared. Certainly no educational researcher concluded that this was a good way to present statistical findings. Nobody wants to read a paper reporting that many different policies could have fairly sizable positive effects, no effect, or a modest negative effect. Papers claiming that popular policies have "insignificant" effects may be politically unwelcome, but editors still prefer the message "nothing works" to the message "everything is uncertain."

By the early 1970s, most social scientists had concluded that if America's goal was to raise student achievement, the policies that most educators favored—smaller classes, better equipment, higher salaries, more extensive teacher training—would not do much good. The bottom line seemed to be that "money doesn't matter." Yet, despite the accumulation of evidence that seemed to point in this direction, neither parents nor educators believed the message. Educators kept asking for more money, legislators kept giving it to them, and the voters mostly went along.

Eventually, educational researchers tired of delivering a message that nobody wanted to hear. Some (including me) turned to other topics. Graduate students in education turned away from quantitative research and began doing qualitative studies, which reduced the risk of finding evidence at odds with

[2] See, for example, the assessments in Mosteller and Moynihan (1971).

their prior beliefs about the world. Quantitative educational research did not disappear, but it was marginalized in most schools of education.

The Impact of Meta-analysis

Quantitative educational researchers took a long time to dig themselves out of this hole. The first step was the invention of meta-analysis, which allowed quantitative researchers to pool results from many different studies in a statistically efficient way. When analysts did this, their view of the world changed drastically. Instead of seeing a world in which most studies yielded "statistically insignificant" coefficients, they saw a world in which most small studies yielded coefficients with the expected sign and in which the average coefficient was large enough to be educationally important.

Gene Glass and his colleagues (1982) showed, for example, that when they pooled results from all the available studies of class size, smaller classes were associated with quite large gains in achievement. This was true despite the fact that most of the original studies had reported an "insignificant" relationship between class size and achievement. Once Glass and his colleagues pooled the data, moreover, the relationship was clearly "significant," in the sense that the confidence interval did not include 0. The trouble with the original studies was that they had been too small to provide reliable information about the size of the effect. Meta-analysis of studies assessing the impact of school desegregation told a similar story, at least for elementary school reading achievement (Cook et al. 1984). So did meta-analysis of studies assessing the impact of most other educational policies (Lipsey and Wilson 1993).

These were all classic cases of Type Two error, where earlier analysts had mistakenly rejected the hypothesis that policies had a positive effect. Students had, of course, learned about Type Two errors for generations. But it was not until the advent of meta-analysis that we began to appreciate both the likelihood and the potential costs of such errors.

Indeed, among some quantitative researchers, meta-analysis led to a dramatic paradigm shift. Instead of assuming that "nothing works," they now began to assume that "everything works." But that too was an oversimplification, for several reasons.

Meta-analysis is feasible only when a policy has been studied many times. Thus, the policy has to remain sufficiently popular over a long enough period to generate numerous studies of its effectiveness. Following Bayesian logic, we would expect policies that have been studied dozens of times to be somewhat more effective than those that have only been studied a few times.

Meta-analysis also highlighted a serious shortcoming of most past educational research. Even when meta-analysts found dozens of studies assessing a particular policy, they seldom found more than a handful of studies that measured the policy's long-term impact. When researchers *had* done long-term follow-ups, moreover, the long-term impact almost always looked smaller than the short-term impact. Among skeptics, therefore, the idea that "nothing works" was gradually replaced by the idea that "everything works in the short run, but benefits usually fade away in the long run."

The idea that achievement gains fade over time may, however, be another artifact of social scientists' statistical conventions. Ever since the invention of IQ tests early in the 20th century, psychometricians have tended to standardize test results. Initially, their goal was to ensure that IQ tests had the same mean and standard deviation at all ages. Standardizing test scores also made it much easier to compare results derived from different tests. But age standardization also obscures a crucial fact about children's cognitive skills, which is that their variance increases with age. If you ask 4-year-olds to do 10 two-digit multiplication problems, their scores are likely to be very similar, because none of the children will be able to do any of the problems. If you ask 14-year-olds to do the same problems, some will get them all right, while others will still get almost all of them wrong. The same logic applies to vocabulary words. The vocabulary of 14-year-olds is more variable than the vocabulary of 4-year-olds.

While this fanning out of academic achievement as children age is well known, meta-analysts usually ignore it. If meta-analysts want to describe the effect of preschool programs at age 4, they will report that those who attended a preschool scored, let us say, 0.30 standard deviations above those who did not attend. If they want to describe the effect five years later, they will report that the gap has shrunk from 0.30 standard deviations to, say, 0.15 standard deviations. They hardly ever ask whether a disparity of 0.15 standard deviations at age 9 is larger or smaller than a disparity of 0.30 standard deviations at age 4.

Meta-analysis has other limitations that curtail its usefulness to policymakers. First, meta-analysts cannot compensate for measurement errors in the studies they pool unless the magnitude of these errors is known. This fact tends to bias estimates of policy impact downward. Second, meta-analysts cannot compensate for the fact that researchers seldom do true experiments, in which randomly selected students are assigned to "treatment" and "control" groups. So-called "natural" experiments, in which students or their parents have a choice about the education they receive, tend to exaggerate the impact of educational policy per se.

Most of the data available to a meta-analyst comes from surveys in which an analyst had identified students affected by some policy of interest, controlled some of the factors correlated with the presence of this policy, and treated any remaining association between the policy and student outcomes as causal. Since analysts can seldom measure all the factors that lead students to have different educational experiences, studies of this kind are likely to suffer from what we once called "omitted variable bias" and now call "selection bias."

Although there is no certain way of eliminating selection bias, better data can often help a lot. The struggle to make quantitative educational research useful to policymakers is critically dependent on these improvements. Fortunately, funders have recognized this. More and more educational surveys now track students over substantial periods of time, measuring both treatments and outcomes on numerous occasions. This strategy provides better estimates of measurement error. It also allows analysts to adjust for the effects of stable unmeasured differences between students. As a result, researchers have been able to generate results that are far more persuasive than those derived a generation ago from the Coleman data or other cross-sectional surveys.

Recent results from analyses of this kind also fit practitioners' expectations better than earlier results did. If this trend continues, researchers may find themselves concluding that conventional wisdom among educators was not as misguided as earlier researchers thought. If researchers can then move on to identifying the most effective strategies for improving achievement, quantitative research may eventually prove quite useful.

The Need for Experiments

Despite all the improvements in survey data and analytic methods, however, there is still one huge problem. Educational researchers used to do experiments in which students were randomly assigned to treatment and control groups. Today, that is extremely rare. To see why the dearth of experiments poses a problem, consider the debate over ability grouping in elementary schools. This is an intensely controversial issue, especially in racially mixed schools. Yet much of the controversy is about facts that would be easy to ascertain using conventional experimental methods. Suppose we ask the following two questions:

1. Do children who have learned little in the past learn more when they are all assigned to heterogeneous classrooms or when classrooms are segregated on the basis of past academic performance?
2. If students who have learned little in the past are assigned to heterogeneous classrooms, do they learn more when the teacher groups them by skill level or when the teacher treats the entire class as a homogenous group?

If students were randomly assigned to different grouping schemes, it would be fairly simple to see which of these schemes was best for which students. But while experiments of this kind are technically feasible, they are no longer done. When Mosteller, Light, and Sachs (1996) surveyed ability grouping experiments done in elementary schools, they found only one study published after 1980. In politics, data that are more than 20 years old carry little weight.

Nor is it clear that such data *should* carry much weight. Mosteller, Light, and Sachs also found only one experiment that dealt with within-classroom grouping, which is more common and more flexible that between-classroom grouping, at least at the elementary school level. They also found that existing experiments dealt exclusively with mathematics. They found no experiments that reported long-term effects. And while they found little evidence that assigning students to classrooms on the basis of past achievement affected the amount learned when all students were taught the same thing, the question that looms largest in today’s debates is what happens when students in “faster” classes cover more material each week, so that they start algebra, geometry, and calculus sooner than they otherwise would.

In theory, it might be possible to answer questions of this kind by comparing school systems that pursue different policies or by following the progress of students who move from one kind of school system to another. In practice, studies of that kind would not be likely to convince skeptics. Experienced school administrators, board members, and legislators know that when a researcher reports results based on complex multivariate statistics, some other researcher will soon come to the opposite conclusion. Complex statistical analyses require dozens of methodological choices that cannot be made by following a generally accepted rule. In a field as politicized as education, the existence of such choices typically guarantees that nonexperimental data will have many possible interpretations.

Those who analyze experiments are also far more likely to agree about what they show. A legislator or a school board member can follow the logic of the Tennessee class size experiment, understand how the results were evaluated, and see why these results mean what researchers say they mean. Of course, legislators do not understand *exactly* what happened. Nor do they understand all the ways in which the experiment may have been contaminated. Least of all, do they understand the limitations of the findings—that the Tennessee results tell us nothing about the benefits of small classes after third grade, for example. But the structure of the argument is still intuitively obvious to almost everyone.

Given these political advantages, why are randomized experiments so rare? The proximate cause is clear: any randomized experiment disrupts school routines, so educators will participate in one only if they think randomization is absolutely crucial to learning something important. To convince educators that experiments are crucial, researchers must be nearly unanimous in supporting the method. Since 1970, such unanimity has vanished. Surprisingly few educational researchers now see experiments as a good way of going about their business. Indeed, few educational researchers have had any experience with randomized experiments.

One reason most researchers are skeptical about experiments is that they seldom really care whether Policy A is better than Policy B. Most researchers care about *why* Policy A is better than Policy B. It takes a whole series of carefully crafted experiments, each of which rules out one or more alternative explanations, to show *why* something happens. And even after dozens of experiments, there is always the possibility that the experimenter has failed to test a plausible alternative to his or her preferred explanation.

Experiments also have a bad reputation because of the way they were once analyzed. The fact that experiments often seemed to show that all kinds of policies did not work has inevitably made educators skeptical about the method. We now know that this pattern was partly traceable to the way we analyzed the data and the way we thought about statistical significance. But even today experiments are politically risky, especially when—as is often the case—they are underfunded and therefore too small to detect small effects.

Still, the shortage of experiments is a huge political problem for anyone who thinks that educational policy should be based on evidence. The fact that the Tennessee class size experiment was funded by the state of Tennessee, not the federal government, is indicative of Washington's failure in this area. When the state that brought us the Scopes Trial funds research more useful than that funded by the federal government, something is terribly wrong in Washington.

Recommendations for Future Data Collection

Although I think we ought to be putting far more resources into experiments, that does not mean we should stop doing surveys. It would be as foolish to stop doing surveys and rely exclusively on experiments as it was to stop doing experiments and rely exclusively on surveys, as we did a generation ago. The two methods are complementary.

If we keep doing surveys, as we surely should, we need to think more carefully than we have about what kinds of data we should collect. Brewer and Goldhaber's (1998) list is a good starting point. Some people may be shocked to hear me say this, because theirs is an economist's list, and I am a sociologist. Nonetheless, I like their list.

Collecting Longitudinal Data

First, longitudinal data *is* definitely better than cross-sectional data. Of course, all researchers think longitudinal data is better as long as they do not have to pay the bill. But I am making a stronger argument, namely that longitudinal data yields more knowledge per dollar than cross-sectional data.

But while longitudinal data is better than cross-sectional data, it is not clear that many years of longitudinal data are better than three or four years. If we want to address problems of measurement error and fadeout, we have to follow students for three or four years. But if surveys keep losing 5 percent of their cases every year because students transfer from one school to another and cannot be followed, at least half the sample will be gone after 12 years. In urban systems where transfers are more common, attrition will be even higher. Furthermore, if a researcher has no findings about older students until her panel of kindergartners has reached college, her funding may dry up in the meantime.

Linking Students to Teachers

Brewer and Goldhaber (1998) are also right about the potential value of linking students to their classroom teachers. If we want to link students to teachers in any meaningful way, however, we also have to collect data in both the fall and the spring, as Meredith Phillips (1998b) argued in her paper for this conference. In addition, we need to ensure that fall and spring testing brackets the school year in a satisfactory way. Testing students in October and April may be convenient, but the school year is more than six months long.

Collecting More Data on Teachers

Brewer and Goldhaber (1998) are also right that collecting more data on teachers is crucial. It is particularly important to gather evidence on the racially charged issue of whether teachers' test scores have a big impact on student achievement, as Ron Ferguson and others have argued (1998). Survey researchers will, of course, have great difficulty testing teachers in today's political environment. That means we need to explore ways of linking our surveys to state records that include teachers' scores on various exams (McLaughlin and Drori 1998).

Domains That Can Be Limited

If we are going to do all this, we also need to identify domains in which we can afford to do less. Surveys planned by committees always have difficulty deciding what to leave out. The committee almost inevitably represents many different interest groups. Indeed, that is usually its main purpose. Such a committee almost inevitably generates a survey instrument that measures many things badly rather than measuring a few things well. As a result, we learn a little about a lot but not much about anything.

That was probably a defensible strategy for the first few national surveys. It is probably not the right strategy for the next century, at least at the secondary level. The broad outlines of what happens in secondary school are now fairly clear. If we are going to collect more data from high school students, we should probably concentrate on one or two topics per survey. At the elementary level, which has been seriously neglected in the past, we may still need several surveys that tell us a little about a lot before we turn to more focused data collection.

Attitudinal Measures

If we have to cut back on certain kinds of survey items, my suggestions would again be similar to Brewer and Goldhaber's (1998). We have collected many attitudinal measures from students over the past generation. I do not think we have learned much as a result. Students' statements about whether they plan to go to college or have a baby out of wedlock do have some predictive power, which means that people's plans are somewhat stable from one year to the next. But that is hardly news. School-to-school differences in students' *average* responses to attitudinal questions may be a bit more useful, since they may tell us something about school "climate." But aggregating behavioral measures probably tells us far more. What we really want to know is whether policies that seek to change attitudes have a long-term effect. Past surveys have seldom tried to determine what schools were doing to change attitudes.

Sampling Strategies

Another crucial issue is whether to sample more students per classroom. Meredith Phillips' (1998a) analysis of the Prospects data indicates that sampling more students per classroom is probably sensible if the researcher wants to study teacher effects, but not for the study of anything else. At least in Prospects, students in the same classroom turn out to be rather similar, so drawing a large sample drawn from a small number of classrooms yields results with large standard errors.

The Question of Representative Samples

I also agree with the Brewer and Goldhaber (1998) that representative samples have been oversold. It is certainly crucial to have representative samples for *some* purposes, but not for *all* purposes. We should think more carefully about the division of labor in data collection. We need occasional national surveys that gather data on large representative samples, perhaps covering a limited number of domains in depth. But we should be able to do a lot of causal modeling with state and local data that are not perfectly representative of any well-defined universe.

The data that state testing programs now collect is of limited interest to most researchers, because states do not gather much background information from students or much program information from schools. But if we could match state data on school achievement to survey data on a subset of students in each school, the result might be ideal for researchers interested in causal modeling. Imagine having the kind of data that Texas collects on individual students linked to the kind of data that NELS collects. Since Texas collects data over a student's entire school career, at least as long as the student stays in Texas, researchers could use such data to answer all kinds of questions they cannot answer now.

Unresolved Problems

I want to close by posing some questions that I think we still need to address if we are to make educational research more relevant to educational policy.

What Can We Learn from Differences between States?

For nearly 100 years, progressive policymakers have claimed that decentralized decision-making gave America an unusual opportunity to learn about the effects of different policies. The states, we were constantly told, were the "laboratories of democracy." In education, however, variations in state policy have taught us surprisingly little. We did learn one thing from the American states' diverse educational experiences in the 20th century, which was that de jure racial segregation had terrible consequences. But we eliminated de jure segregation a generation ago.

Since 1970, the 50 American states have continued to pursue 50 different sets of policies. As far as I can tell, we have learned nothing from this diversity. In part, this is because we have not collected good information about what policies states were really pursuing. In part it is because we have not collected good information about how educational outcomes differed from state to state. Primarily, however, our failure to learn from states' experiences reflects the fact that learning from such experiences would require detailed data linking year-to-year changes in educational policy to changes in subsequent outcomes. We simply do not have such information.

The papers that David Grissmer (1998) and Steve Raudenbush (1998) presented this morning suggest that NAEP may now provide state-level outcome measures that can offer useful policy guidance. Grissmer certainly drew some interesting lessons from North Carolina and Texas. But Raudenbush found that a handful of variables explain almost all state-to-state differences in student achievement. If that is true, these laboratories of democracy may be rather like a child's chemistry set that includes only a handful of different chemicals. This does not imply that educational policy is unimportant. But it does imply that states may be so internally heterogeneous in the policies their schools pursue that state means on the policy variables that matter are relatively similar. This issue requires further exploration.

How Should We Define and Measure School Achievement?

Educators try to teach specific skills and information. If we want to make research useful to educators, we have to measure the skills and information that educators try to teach. We also have to measure skills and information on scales that allow us to say how much students have learned between one period and the next. That requires two major changes in the way educational researchers go about their business.

First, we probably have to stop equating the quality of a test with its reliability. The way to get high reliability is to choose items that are highly correlated with one another. But in a world with diverse schools, teachers, and curriculums, tests with high inter-item correlations almost always end up measuring general ability, not the specific skills and information that educators in particular places have tried to impart.

The other far-reaching change we will have to make is to stop standardizing tests to predetermined means and variances. I do not think we will ever make much progress in measuring learning if we keep thinking about achievement in exclusively relative terms, as we mostly have for the past 100 years.

To see how misleading relative rankings can be, consider the Tennessee class size experiment. Tennessee assigned children to either large or small classes from kindergarten through the end of third grade. The children assigned to smaller classes did better at the end of kindergarten. They preserved, but did not widen, their advantage over the next three years. Many people have been puzzled by the fact that smaller classes did not seem to yield any further benefit after the first year of the experiment. Skeptics like Eric Hanushek (1999) have interpreted this finding as evidence that only the first year in a small class yields measurable benefits. But, as Jeremy Finn pointed out in the discussion today, this picture reflects the fact that those who analyzed the Tennessee data reported treatment effects at different ages as a percentage of the standard deviation at that age.[3] When one looks at unstandardized treatment effects, they increase as time goes on.

How Can Educational Research Serve Non-policy Goals?

Up to this point I have discussed educational research as if its only legitimate purpose was to improve educational policy by helping policymakers decide whether small classes are worth the extra cost, whether to group students on the basis of their past achievement, and so on. But educational research has another function as well. It tells us something about how we are doing as a people and as a nation. It helps us compare our performance both to our ideals and to the performance of other democratic societies. As a result, it plays a significant role in our judgments about whether we live in a just or an unjust society, whether opportunity is more equal in our society than in other societies, and whether things were really better a generation ago, when most of us were growing up. These are not policy questions in any ordinary sense. But how we answer these questions has an important impact on how we think about ourselves and what policies we favor or oppose.

If American students learn less math than Japanese students, for example, it is not at all clear what policy implications this fact should have. But we still want to know the fact. Similarly, when the black-white test score gap falls dramatically, as it apparently did during the 1980s, this brute fact does not tell us anything about *why* the gap fell or what policies might further reduce it in the future.[4] But even when facts of this kind have no clear policy implications, they tell us something about what is happening to our country that we should—and do—care about.

If I am right in claiming that one major goal of educational research is to tell ordinary citizens how well their country is doing in various domains, we need to report such information in a form that most citizens can understand. I have already argued that comprehensibility is a major argument for preferring randomized experiments to multivariate statistics. The need for comprehensi-

[3] Comments from Jeremy Finn during the discussion period of the seminar, November 9, 1998.

[4] See Grismer, Flanagan, and Williamson (1998).

bility should also play a central role in the way we report data on states and school districts. One example must suffice. Researchers often report mean scores for states or school districts that have been adjusted statistically to eliminate the effects of demographic differences between states or districts. This kind of adjustment poses many statistical problems. But its most serious defect may be that it is very hard to explain. Thus, wherever possible, it is probably better to present separate means for different kinds of students: those whose parents have had different amounts of schooling, for example, or those from different racial and ethnic groups. Readers can then see for themselves how each group fares in different states.

Can We Predict the Future?

If anyone had asked me in the late 1960s what educational researchers would learn over the next 15 years, I would have predicted far more progress than actually occurred. It is stunning how little progress we really made between the late 1960s and the early 1980s. But if anyone had repeated the question in the early 1980s, I would again have been wrong. It would never have occurred to me that after 15 years of spinning our wheels, we were about to make progress at an unprecedented rate. Yet that is what happened. The papers presented at this seminar are, I think, much better than anyone would have predicted 15 years ago.

We should all reflect on this history and try to use it to identify the preconditions for intellectual progress. Why, after an apparently promising start in the 1960s, did we accomplish so little in the next 15 years? Why did things go so much better over the next 15 years? I have my own hunches. Meta-analysis forced us to rethink our approach to significance tests. NCES began collecting better data. Econometricians pushed us to adopt more rigorous standards of proof. But these are speculations. A careful intellectual history of quantitative educational research remains a task for the future.

References

Brewer, D. J., and Goldhaber, D. D. (1998). Improving Longitudinal Data on Student Achievement: Some Lessons from Recent Research Using NELS:88. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 169–188). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Cook, T., Armor, D., Crain, R., Miller, N., Stephan, W., Walberg, H., and Wortman, P. (1984). *School Desegregation and Black Achievement*. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Institute of Education.

Ferguson, R. F., and Brown, J. (1998). Certification Test Scores, Teacher Quality, and Student Achievement. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 133–156). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Glass, G., Cahen, L., Smith, M. L., and Filby, N. (1982). *School Class Size*. Thousand Oaks, CA: Sage Publications.

Grissmer, D. W., and Flanagan, A. (1999). Moving Educational Research toward Scientific Consensus. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 43–90). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Grissmer, D.W., Flanagan, A., and Williamson, S. (1998). Why Did the Black-White Test Score Gap Narrow in the 1970s and 1980s? In C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 182–228). Washington, DC: Brookings Institution Press.

Hanushek, E. A. (1999). The Evidence on Class Size. In S. E. Mayer and P. E. Peterson (Eds.) *Earning and Learning: How Schools Matter*. Washington, DC: Brookings Institution Press.

Jencks, C. (1971). The Coleman Report and the Conventional Wisdom. In F. Mosteller and D. P. Moynihan (Eds.), *On Equality of Educational Opportunity* (pp. 69–115). New York: Random House.

Lipsey, M., and Wilson, D. (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-analysis. *American Psychologist 48*(12): 1181–1209.

McLaughlin, D., and Drori, G. (1999). School-level Correlates of Reading and Mathematics Achievement in Public Schools. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 189–236). U.S. Department of Education. National Center for Education Statistics. Washington, DC:

Mosteller, F., Light, R., and Sachs, J. (1996). Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size. *Harvard Educational Review 66*: 797–842.

Mosteller, F., and Moynihan, D. P. (Eds.) (1971). *On Equality of Educational Opportunity*. New York: Random House.

Phillips, M. (1998). *Do African-American and Latino Children Learn More in Predominantly White Schools?* Unpublished manuscript, School of Public Policy and Social Research, University of California, Los Angeles.

Phillips, M. (1999). Understanding Ethnic Differences in Academic Achievement: Empirical Lessons from National Data. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 103–132). (U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Raudenbush, S. W. (1999). Synthesizing Results from the NAEP Trial State Assessment. In D. W. Grissmer and J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement* (NCES 2000–050) (pp. 3–42). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Improving Research and Data Collection on Student Achievement[1]

Brenda J. Turnbull, Policy Study Associates; David W. Grissmer, RAND; and J. Michael Ross, National Center for Education Statistics

This seminar brought together a diverse group of people linked by their motivation to understand and improve student achievement. Participants in the seminar included researchers whose work has focused on analyzing student achievement, policymakers instrumental in designing policies to improve student performance, and government officials who design and manage the collection of major data sets used by researchers and policymakers. The seminar sought to lower the inevitable communication barriers existing within this community: between researchers and policymakers, between those designing and collecting data and those who use it; and among researchers from different disciplines analyzing different data sets with different models and estimation techniques. From the divergent perspectives, we sought to identify directions for future research and data collections, and perhaps a common conceptual framework encompassing research and data collection on achievement.

Here we summarize the recommendations made by participants for more sophisticated data collection strategies, new directions for future research, and collaborative forums for communication of research results. These recommendations fall into two broad categories. The first category includes a variety of smaller incremental changes focusing on improving nonexperimental results, while the second category includes more radical departures from current directions in federal statistical agencies. Finally, we focus on a different topic of discussion at the seminar—communication of results—and provide some concluding remarks on future directions.

[1] The authors are grateful to Martin Orland, John Ralph, Joseph Conaty, and Daniel Kasprzyk for an early discussion of the implications that could emerge from this seminar. However, the opinions expressed herein are those of the authors; and no endorsement by the National Center for Education Statistics, the Office of Educational Research and Improvement, or RAND should be inferred.

Improving Nonexperimental Research and Data Collection

Improving longitudinal surveys was the focus of many recommendations. The National Center for Education Statistics (NCES) and the Department of Education fund several key longitudinal surveys, such as the National Education Longitudinal Survey (NELS:88) and the Early Childhood Longitudinal Studies (ECLS). Since achievement tests are included in their designs, during the last 20 years these surveys have supported a significant amount of research on achievement. Seminar participants could—and did—disagree over the advantages of specific analytic procedures, but none questioned the essential value of having and expanding the collection of longitudinal data. The recommendations for improving existing longitudinal studies centered on the inevitable competing priorities for specific topics and survey questions, the frequency of data collection, more stratified sampling plans, and the importance of prekindergarten baseline measures, as well as a more structured process for the design of all surveys. Dominic Brewer and Dan Goldhaber presented a list of specific recommendations that also received support from Christopher Jencks; indeed, most participants contributed recommendations in this area.

One key issue in survey design was highlighted by an exchange from the floor. Jeremy Finn suggested that a survey should focus on only six to twelve well-specified constructs. Jencks countered that a government agency cannot make such a draconian selection, because it is answerable to many constituencies. Jencks did suggest, though, that NCES could take a retrospective look at the actual uses made of particular variables and particular items, saying that this kind of analysis would support the selective deletion of less productive items from repeated surveys. This inductive procedure would be quite different from the deductive one proposed by Finn, where a clear and bounded conceptual framework would drive the construction of survey items. These remarks reflect the reality that a large-scale survey qualifies both as research and as a political undertaking, because the data ultimately are used for many purposes. However, scientific constructs can improve the items devoted to research issues and at the same time perhaps constrain the usually high demand for items devoted to nonresearch issues.

An allied issue is the discontinuation of past survey items in order to include new items. From time to time during the day, a participant would suggest ways of trimming the length of surveys by deleting specific items.

Generally, another participant would swiftly object, arguing for the theoretical or practical significance of those items. An example was the suggestion by Brewer and Goldhaber that "school climate" measures could be substantially reduced, which was followed by a rejoinder from Valerie Lee that these measures are integral to an assessment of school quality. Such disagreements over items partly reflect the still unresolved disciplinary preferences for the importance of certain variables and modes of analysis. These disagreements suggest that design teams should include researchers and scholars from different disciplines insofar as they can engage in productive dialogue over these issues from their different perspectives.

Suggestions were also made concerning the frequency and timing of collecting longitudinal data. Collecting data each year rather than every two years when following a cohort through school was mentioned, but the increased costs or trade-offs with other survey design parameters would certainly need to be considered. Meredith Phillips made a compelling case for shifting the annual testing of students to testing in both fall and spring. Although burdensome in implementation, this change would permit analysis and comparison of students' learning trajectories during the regular school year and the summer. The importance of this issue was demonstrated in Phillips' analysis, which confirms that the test score gap between black and white students widened during the summer time period. Her recommendation was echoed by Smith, Jencks, and, from the floor, Adam Gamoran.

Recommendations were also made with respect to sampling strategies relevant to classroom-within-school effects. Brewer and Goldhaber and Lee supported nested samples with more students per classroom. Lee suggested that more students per class could be efficiently traded off for fewer schools in longitudinal samples. In addition, Brewer and Goldhaber, along with Ron Ferguson and Jordana Brown, suggested sampling more teachers per school. Realistically, these changes would increase costs and burden, but their endorsement by researchers does suggest that these trade-offs should at least be given serious consideration in the design of future samples.

Several presenters emphasized the desirability—as well as the challenges—of finding out "what teachers know." Stephen Raudenbush mentioned the importance of measuring teachers' subject matter preparation and content knowledge in their assigned teaching fields. For Ferguson and Brown, teachers' scores on tests such as the ACT are also important data, although Ferguson said that NCES should “probably not” try to collect such data itself, but should instead play a facilitating and convening role with private sector organizations that already collect such data. Brewer and Goldhaber suggested such possible measures as administering a written assessment to teachers or asking principals to assess teacher quality.

Finally, David Grissmer and Ann Flanagan suggested that longitudinal surveys need to begin at school entry in order to capture the variables needed for accurate estimation of both short- and long-term effects in student learning. They suggest that production function methodologies that utilize only a previous test score as a proxy for earlier schooling variables are challenged by the results of the Tennessee experiment. They observe that these results suggest a multiyear effect from class size reductions, and these multiyear effects cannot be controlled for by a single-year previous test score. Therefore, survey resources should be shifted to earlier years, and the value of surveys started at later grades would be diminished.

Improving NAEP Data

Data from the National Assessment of Educational Progress (NAEP) provide the only achievement data with representative samples of U.S. students. NAEP has been administered to samples of students at ages 9, 13, and 17 since 1969 in several subjects. While the trend results from these data have been often used and cited in many studies, research explaining the trends is more recent. The NAEP data remain the most reliable information for assessing the changing gap between minority and white students, and also for facilitating inquiries into the question of whether educational and social policies directed toward minority and low-income families and students have raised student achievement test scores. The expansion of NAEP, since 1990, to state samples also provides the only comparative achievement scores across states with representative samples within each state. Several recommendations for improving the NAEP data emerge from the research presented in the seminar.

Grissmer and Flanagan observe that the most serious weakness of NAEP can be found in the student-reported family characteristics. They suggest implementing a simple parental survey as one means of collecting better data. However, they also suggest testing empirically whether student-reported surrogate measures such as “books in the home” can provide adequate substitutes for parent-collected data. They also suggest using state-level Census data to supplement NAEP family data.

However, any significant addition to the NAEP data collection such as a parent survey may have significant opportunity costs. Sylvia Johnson and Gary Phillips mentioned the current press to speed up and simplify NAEP, pointing out that this works against the more complex data collection that would permit us, in Johnson's words, to "better understand the whys and hows of improving student achievement."

Johnson, as well as Grissmer and Flanagan, made the more radical suggestion of changing the sampling plan from school samples to school district samples. Several advantages might accrue from a district-based sample, although the more complex sampling would increase costs. A district-based sample would allow comparing and explaining achievement differences across major urban school systems relative to smaller, more homogeneous suburban districts. Urban school systems encompass a significant part of the nation's education problems, but we currently have no adequate comparative measures of performance across these urban systems. Johnson observed that the change to school district sampling could help spur improvement in learning because it would heighten public scrutiny of districts' results. However, if costs were kept constant, such sampling would probably mean fewer students per school, allowing more variance in measurement of school characteristics.

Recommendations for Changing Directions

Pointing to the near universal credibility enjoyed by Tennessee's experiment with reduced class size, several presenters called for more experiments (randomized field trials) to evaluate education programs. Deputy Secretary Marshall Smith characterized the policy impact of the Tennessee experiment as "instantaneous" and "incredibly powerful in Congress." Compared with conventional surveys, he said, such experimental trials provide more robust and probably more valid estimates of program effects. Also emphasizing the power of experiments to communicate, Jencks credited experiments with producing results that are easy to understand, saying "The structure of the argument is intuitively obvious." Grissmer and Flanagan argued that research consensus is more likely to emerge when a well-designed, -implemented, and -analyzed experiment has taken place and when the analysis can show little sensitivity to the inevitable deviations from ideal design specifications. Although nonexperimental studies can be filtered through the scrutiny of meta-analysis and expert panels, they observed that this process often does not lead to consensus.

Recognizing the high costs and the limits of experiments, several speakers also addressed strategies for accumulating evidence over time through a series of inter-related investigations. Smith cited the studies related to the acquisition of reading skills supported by the National Institute of Child Health and Human Development (NICHD), which he described as "not a single experiment, but a series of interlocking studies testing hypotheses about reading." These studies enabled NICHD to "tell a very coherent story about reading, with immense power" in policy circles, and thus they have paved the way for the appropriations needed to further experimental studies. Jencks, in answer to a question from the floor, endorsed the idea of embedding small experiments within more conventional survey designs.

Deputy Secretary of Education Smith proposed several topics which would benefit from experimental investigations, each representing an area in which there is a theoretical and an evidentiary base upon which to design randomized trials, as well as strong policy interest:

- ***After-school and summer programs.*** The rationale for offering extended learning time is clear, but little or no good evidence is yet available on the effects of well-designed programs that provide students with a safe environment and adult tutoring beyond the typical school day and year.
- ***Training in parenting skills.*** Such training could be combined with adult education in an experimental trial.
- ***Education-focused preschool programs***. Evidence now indicates that children benefit from acquiring particular skills and concepts before they enter school, such as knowing the alphabet, knowing that one reads from left to right, knowing the concepts of *before* and *after*, *up* and *down*, and the like. The effects of such instruction could be studied experimentally.
- ***Research-based interventions in school reform.*** David Cohen of the University of Michigan is launching a major study of schoolwide reforms. His methods involve detailed survey and observational methods rather than an experimental test of the policy that encourages schools to implement well-specified programs under the guidance of outside

experts. Despite the current enthusiasm for schoolwide reforms, they have not been well studied in either education or business.

What could not be explored in much depth in the discussions, however, was the place that experiments might occupy in an overall portfolio of support for research and how decisions would be made concerning such a portfolio of experiments. Jencks noted that the Tennessee experiment was launched by the state of Tennessee without federal help, and the state of Wisconsin has also initiated a quasi-experiment with student-teacher ratios. Both NCES and the Office of Educational Research and Improvement (OERI) could face uphill struggles, for somewhat different reasons, if they were to try to sponsor experiments. For a statistical agency to take the lead in initiating an experiment would be unusual and might jeopardize the stability of its role and mission, at least according to some observers. A well-funded research agency can include experiments in its portfolio, but OERI currently lacks the discretionary resources to launch such costly investigations. This dilemma suggests the need for an innovative federal-state partnership in the organization and implementation of an experimental agenda.

In the last 10 years, state testing programs have become the major source of achievement data in the nation. Virtually every state is now committed to more frequent testing of its students statewide in a variety of subjects across a variety of grades. The largest representative sample collected in national achievement data is approximately 25,000 students, and the largest state samples collected by NCES (NAEP Trial State Assessment) usually test around 2,500 students per state. Samples for state-administered tests include nearly all students with the exception of certain IEP (Individual Education Program) and LEP (Limited English Proficient) students who are excluded, so sample sizes can be over one million students in several states. National tests are typically given every four years, while state testing often occurs on an annual basis within certain grade ranges. Moreover, in some states, individual student scores can be tracked across grades and linked to specific teachers, thereby allowing even richer longitudinal and contextual analysis.

This explosion of achievement testing in many states suggests a new direction for federal data collection effort; namely, using state achievement data as a platform for research and experimentation. The McLaughlin and Drori study provides an example of linking this state assessment data at the school level with federal data such as the Schools and Staffing Survey (SASS). State achievement data can also be supplemented with state teacher data, more detailed information on resources, and facility information. The Ferguson and Brown analysis illustrates the power of linking these data with teacher characteristics and other data available in Texas. Their paper drew from state databases that permit analysis of teacher test performance in relation to student performance at the district level in Texas. Ferguson recommended more such work, but observed that NCES would not have to administer tests to teachers in order to obtain data. Instead, Ferguson suggested that the federal government play a more active role of leadership in convening and coordinating research efforts with organizations such as Educational Testing Service (ETS) and perhaps with state agencies through the exchange of data with appropriate protections for individual confidentiality.

Yet the potential for a richer "universe" as sources of data may lie in evaluating planned interventions and experimentation. Many states already have many ongoing education reforms that could be better evaluated with these data sets. However, a federal leadership role would encourage both random assignment of schools or districts by states in the initial phase of program implementation and then the funding of high quality program evaluations conducted by national experts.

Both controlled field trials and longitudinal studies are usually seen as intrinsically complex and costly endeavors. However, such research can range from small to large scale. Departing from the general endorsement of loading more complexity onto one large multipurpose study, Robert Hauser suggested that an alternative would be the more frequent initiation of smaller longitudinal studies, such as the Wisconsin Longitudinal Study. These would differ from the "larger, one-time-only or once-per-decade surveys" that have been customary in education. He elaborated:

> We ought to be initiating cohort surveys close to birth every year—or every other year—as a means of improving our "who, what, when?" understanding. Such surveys should be stratified by ethnic origin, differentially sampled. And they should provide opportunity for experimentation with alternative test (and questionnaire) content and observational designs, as well as for core content stable enough to permit aggregation of findings across cohorts to yield greater statistical power.

Such small-scale but in-depth studies can lead to better model specifications in larger studies, even if they are not totally representative of the entire population. Hauser argues that it is hard to successfully model achievement in large samples, until we understand more precisely the development of individual students in different contexts. Such smaller scale studies, if focused on testing the assumptions usually made in larger scale research studies, could probably contribute significantly to the development of more coherent research findings.

Small-scale experiments are also possible, given that achievement data are usually already collected across many different states. Different types of interventions are occurring in schools, and early development models may be easily turned toward at least a quasi-experimentation orientation in the early stages of implementation. Universities in each state could become centers for initiating intervention research and experimentation that builds on current assessment data. In discussing research on existing databases, Ferguson and Brown, based upon their experience in Texas and Alabama, emphasized the value of working with researchers who are based in a state and have a long-term career interest in working with that state's database. Again, the federal role would be to convene researchers around a shared knowledge-building agenda, rather than simply to supply data.

Communicating Research Results

The seminar's recurring emphasis on the communication of new findings often challenged conventional assumptions and served as a reminder that statistical and explanatory presentations are an integral part of more sophisticated methods of data collection and advances in data analysis techniques. Further, decisions about the specific form of a presentation can make a difference in the persuasive power and ultimate value of data and research for policymakers.

Smith described the power of the Tennessee experiment to influence policymakers—a result Jencks attributed to the transparency of experimental results. Jencks noted the quite different implications that are often communicated by displaying aggregate achievement data versus the same data when it is disaggregated into different demographic groups. For instance, the display of aggregate NAEP trends has often been used to imply that no gains have occurred in achievement over the last 25 years, while the display of minority trends shows significant gains in achievement, at least up to 1988.

Communicating nonexperimental results is more difficult due to the nontransparency of the analysis. However, Raudenbush's paper provides a compelling example of how a display of nonexperimental results illustrates the varying factors related to achievement differences between states. The graphical display of achievement differences across states illustrates the differential resources and opportunities across states. In both previous and the current work, Raudenbush has also advocated "value-added" models of achievement when evaluation of schools or teachers is involved. Separating the impact of family and social capital from specific schooling effects is a part of being able to effectively partition the components of achievement differences and effectively communicate results.

Some other participants also addressed the topic of communication. Grissmer advocated more emphasis on support for professional consensus activities through conferences, National Research Council panels, and the preparation of edited books focused on specific topics.[2] Support for consensus panels, which are common in health research, was expressed, the intent being to provide a more formal basis to study, form consensus, and communicate important research conclusions.

Concluding Remarks

Reviewing the seminar papers in their totality, we are encouraged by the large and increasing amount of achievement data being collected—some longitudinal, some cross-sectional—and we are quite optimistic that some persistent and perplexing research questions are now empirically answerable. Besides the data collected by NAEP and the combination of main and State NAEP, it is now possible to determine the relative variations between states and between schools within states. In addition, the research is now beginning to examine in more detail the resource differences associated with these outcome measures. In addition, some states have collected longitudinal databases at the student level for all schools and districts, therefore permitting researchers to measure grade-specific changes in student performance over their years in school. Many of these new databases also allow the statistical examination of multilevel factors (such as school- and district-level factors) and social context factors, where the performance of certain types of students may be quite different depending

[2] Examples of recent books include the following: Ladd 1996; Burtless 1996; Jencks and Phillips 1998; Ladd, Chalk, and Hansen 1999; and Ladd, Chalk, and Hansen 2000.

upon the social and racial composition of the school. On the horizon, data will be available for a new longitudinal cohort of students beginning in kindergarten and extending through fifth grade, which are coupled with parent, teacher, and administrator surveys; these new data sets should fill in many gaps in our knowledge concerning instructional practices and the relative importance of family and out-of-school social factors. Finally, we are optimistic that, in the near future, new governance arrangements will expand opportunities for state and federal researchers to link and analyze new databases, thereby generating more refined statistical estimates of these factors.

The accurate and valid measurement of student achievement and performance will increase in importance over time. Not only is “testing” becoming more “high stakes” in some states, but research on these issues is also becoming “high stakes.” For better or worse, student achievement is increasingly being used to measure the effectiveness of schools in some states, as well as the overall effectiveness of the resources committed to education change. The federal government, state legislatures, and district policymakers increasingly utilize the findings from this research to guide and justify their policies; and research findings are commonly utilized in equity and adequacy lawsuits being pursued in many states. Likewise, the “quality of education” ranks near the top of voters’ concerns in national, state, and local elections; and, unfortunately, candidates for public office are tempted to utilize achievement test scores as evidence for the success or failure of new reform policies. Significant increases in educational spending are a common part of the agenda of both political parties, and these increases are now often linked to accountability for producing higher achievement.

Although the findings from education research are being cited and utilized more frequently by policymakers, a wide range of inconsistent, and at times contradictory, research results are often put forth. Researchers have many different explanations for why findings are inconsistent. One explanation attributes the inconsistency to the actual difference in the effects across different contexts. This explanation trusts the modeling process to accurately estimate a “real” effect, but expects the results to differ due to the different contexts. For instance, two measurements of the effects of per pupil expenditure might differ because the money may be allocated more effectively in some cases, and the inconsistency is interpreted as reflecting a public school system that lacks incentives to utilize money well.

A second explanation attributes the inconsistent results to flaws in the modeling process, and not to “real” differences in effects. It assumes that the inconsistent findings reflect imprecise modeling methods that do not fully reflect the complexity of the process being measured. Previous models are rarely comparable, because different data sets are used with varying quality of data or different model specifications or different estimation methods. In this case, widely varied findings are interpreted as inconclusive, and conclusions are not drawn from these models about the efficiency or effectiveness of school systems.

It is difficult to determine how much of the inconsistency is due to imprecision in modeling versus real contextual variation in the actual effect when we rely primarily on nonexperimental measurements, but several directions for research can be helpful. The first is to use similar model specifications and estimation across a variety of large data sets to eliminate certain sources of variance—different specifications and estimation methods.[3] A second direction is to utilize data sets with the most complete sets of variables to explore the sensitivity of variable coefficients to less complete variable sets.[4] (For instance, many databases contain few family background variables, often just race-ethnicity and a measure of income obtained at one point in time, such as qualifying for free and reduced price lunch. Data sets containing a much richer set of family resource variables can be used to explore which schooling variables are affected by less complete sets of family variables and how much sensitivity exists.) A third direction is for literature reviews to address directly the question of why many findings differ across studies rather than simply note that findings are widely dispersed.[5] A fourth direction is to use different estimation techniques across the same data sets with similar variables to determine how sensitive these measures are to different estimation methods. A fifth direction is to empirically test directly some of the major assumptions used in

[3] One example of this is Hedges and Nowell (1998). They fit models with similar family variables across a variety of data sets that have achievement and family variables.

[4] An example of this is Goldhaber and Brewer (1997).

[5] A recent example of this is Krueger (1999). Krueger takes the findings from each study used in Hanushek, Rivkin, and Taylor (1996) and tests whether the conclusions are sensitive to the criteria used for measurement inclusion and level of aggregation. Other examples include Greenwald, Hedges, and Laine (1996) and Hanushek (1996), where the sensitivity of results to output measures, specification, time of publication, school level, and level of aggregation are tested. Grissmer et al. (forthcoming) also try to explain the pattern of different results of previous measurements by level of aggregation.

nonexperimental analysis.[6] There are examples in the literature of each of these five types of analysis, but there is not a coherent effort to improve our nonexperimental analysis based upon these studies.

Focusing on why nonexperimental measurements produce different findings and discovering the underlying hypothesis that explains the pattern of variance can be helpful, but it is doubtful whether research consensus can emerge from this process alone. There are simply too many different possibilities that might explain why nonexperimental measurements differ and current data sets have significant limitations in their ability to support such analysis. However, two complementary directions—experimentation and micro-process theory building—may eventually provide decisive evidence that can explain the variance in nonexperimental analysis and lead to a convergence of evidence from which scientific consensus can emerge. Well-designed, -implemented, and -analyzed experimental data that show little sensitivity to the inevitable deviations from ideal design can serve two purposes. First, more reliable measurements can serve as benchmarks for evaluating results from nonexperimental models. Second, experimental data can serve to calibrate nonexperimental models by identifying specifications and estimations that can predict experimental results. Thus, experimental data can contribute to our understanding of why inconsistent findings exist in nonexperimental measurements and why convergence between experimental and nonexperimental evidence may eventually emerge.

Still, finding consistency across experimental and nonexperimental evidence leaves out one important element from which scientific consensus develops. The power of theories is that they can successfully explain past empirical results and predict future empirical results and ultimately provide the authority from which scientific consensus emerges. Theories that successfully predict how teachers and students will change behavior in smaller classrooms, and how that changed behavior leads to higher achievement, and why there are different behavior changes in classes with high and low SES begin to generate the authority for consensus. The ultimate test, however, is for the theory to generate new constructs and operational definitions and subsequently to predict their effects on achievement outcomes.

[6] An example of this is Heckmaan, Layne-Farrar, and Todd (1996).

Theory building does not receive much emphasis in education research, partly because it is difficult terrain to negotiate. Theory building is necessarily multidisciplinary because the building blocks come from psychology, child development, cognitive science, economics, and sociology. Further, it involves human behavior as well as developmental processes. However, theory building can begin with much simpler goals, such as predicting how achievement changes when resources or teachers change by linking changes in classroom and home behavior when resources change. Successful theories are not only an ultimate goal of good scientific research, but also can be very effective communication tools since they tell "stories" of why resource changes affect achievement.[7]

Theory development requires a coordinated research agenda and more comprehensive data collections. The Tennessee class size experiment, with its associated collection of classroom data, was a good beginning for the 1980s. But this experiment could have collected better data aimed at explaining why achievement was sustained after students left the smaller classes in later years in school. What behavior or developmental pattern changed to sustain these results through eighth grade, and why did those students in small classes for only one to two years not sustain gains as compared to those in small classes for three to four years? More detailed classroom and home observations, more information concerning peer and family relationships, and eventually brain processing patterns in early and later grades for students in small and large classes can help explain such effects, and provide the basis for theory building.

Reflecting on the presentations, discussions during the seminar, and the revised papers in this book, we are hopeful that the deliberations have enabled federal researchers and other policymakers to take stock of innovative research, and especially of achievement research, as areas of empirical inquiry. This process of rethinking basic assumptions should facilitate new understandings of where education research has been in the past and, more importantly, where new challenging research opportunities may be presenting themselves in the future. Thus, our efforts to ensure richer quality of information from our data sources, improved methods of empirical inquiry, and more informative theory building should be enhanced by an occasion when different groups come together to exchange ideas and present written summaries of their findings. It is,

[7] See *Educational Evaluation and Policy Analysis 20*(2) (1999, summer) on class size for an example of linking research together to support theory building.

for example, gratifying to realize the abundance of data now available, compared to the skimpy information prior to the publication of “the Coleman report” in 1966. Further, it is encouraging to recognize the expanding applications to education policy that are the results of utilizing federal and state data collections. For the diverse research community encompassed in the seminar, intellectual challenges remain: fine-tuning existing data collection strategies, exploring linkages and connections between federal and state data sources, improvements in specifications within statistical models, and building relevant and useful theories of education processes. Thoughtful consideration of selected recommendations contained in this report should lead to a more coherent and productive research agenda for federal statistical agencies and state research organizations; to promising partnerships between these researchers and independent entities such as ETS; and eventually to an expansion of knowledge that facilitates and promotes student learning in schools.

References

Burtless, G. (1996). *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington, DC: The Brookings Institution Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., York, R. L. (1966). *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Goldhaber, D. D., and Brewer, D. J. (1997) Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *Journal of Human Resources 32*(3): 505–523.

Greenwald, R., Hedges, L. V., and Laine, R. D. (1996, fall). The Effect of School Resources on Student Achievement. *Review of Educational Research 66*(3): 361–396

Grissmer, D. W., Flanagan, A., Kawata, J., and Williamson, S. (forthcoming). *Improving Student Achievement: State Policies That Make a Difference.* Santa Monica, CA: RAND.

Grissmer, D. W. (1999, summer). Assessing the Evidence on Class Size: Policy Implications and Future Research Agenda. *Educational Evaluation and Policy Analysis 20*(2).

Hanushek, E. A., Rivkin, S. G., and Taylor, L. L. (1996, fall). Aggregation and the Estimated Effects of School Resources. *The Review of Economics and Statistics*, pp. 611–627.

Heckman, J., Layne-Farrar, A., and Todd, P. (1996). Does Measured School Quality Really Matter? In H. F. Ladd (Ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (pp. 192–289). Washington, DC: The Brookings Institution Press.

Hedges, L. V., and Nowell, A. (1998). Group Differences in Mental Test Scores: Mean Differences, Variability, and Talent. In C. Jencks and M. Phillips (Eds.) *The Black-White Test Score Gap.* Washington, DC: The Brookings Institution Press.

Jencks, C. and Phillips, M. (1998). *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Krueger, A. B. (1999). *An Economist Looks at Class Size Research.* Working Paper. Princeton, NJ: Princeton University.

Ladd, H. F. (Ed.). (1996). *Holding Schools Accountable.* Washington, D.C.: The Brookings Institution Press.

Ladd, H. F., Chalk, R., and Hansen, J. S. (1999). *Equity and Adequacy in Educational Finance: Issues and Perspectives.* Washington, DC: National Academy Press.

Ladd, H. F., Chalk, R., and Hansen, J. S. (2000). *Making Money Matter.* Washington, DC: National Academy Press.

Section V.
Appendix

Seminar Attendees 319

Seminar Attendees *(as of November 1998)*

Nancy Allen
Educational Testing Service
Rosedale Road, Mailstop 02T
Princeton, NJ 08541
E-mail: nallen@ets.org

Judith Anderson
U.S. Department of Education
National Institute on Student
Achievement, Curriculum, and
Assessment
555 New Jersey Ave., NW
Room 517
Washington, DC 20208
E-mail: judith_anderson@ed.gov

Pat Anderson
District of Columbia Public
Schools
Department of Educational
Accountability
825 N. Capitol Street, NE
Room 8053B
Washington, DC 20002

Mark Berends
RAND
1333 H Street, NW
Washington, DC 20005–4707
E-mail: mark_berends@rand.org

Sue Betka
U.S. Department of Education
Office of Educational Research
and Improvement
555 New Jersey Ave., NW
Room 600G
Washington, DC 20208–5649
E-mail: sue_betka@ed.gov

Eve Bither
U.S. Department of Education
National Education Research,
Policy &
Priorities Board
555 New Jersey Ave., NW
Room 101
Washington, DC 20208

Rolf Blank
Council of Chief State School
Officers
One Massachusetts Ave., NW
Suite 700
Washington, DC 20001–1431
E-mail: rolfb@ccsso.org

George Bohrnstedt
American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302–1113
E-mail: gbohrnstedt@air-ca.org

Geoffrey D. Borman
The Johns Hopkins University
3003 North Charles Street
Suite 200
Baltimore, MD 21218
E-mail: gborman@csos.jhu.edu

Anne Bouie
U.S. Department of Education
Office of Educational Research
and Improvement
555 New Jersey Ave., NW
Suite 611c
Washington, DC 20208
E-mail: anne_bouie@ed.gov

Mary Lyn Bourque
National Assessment Governing
Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002
E-mail:
mary_lyn_bourque@ed.gov

Dominic Brewer
RAND
1333 H Street, NW
Washington, DC 20005–4707
E-mail:
dominic_brewer@rand.org

Jordana Brown
John F. Kennedy School of
Government
Malcolm Wiener Center for
Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail: jordana_brown/
fs.ksg@ksg.harvard.edu

Janis Brown
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 318
Washington, DC 20208
E-mail: janis_brown@ed.gov

Peggy G. Carr
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 308
Washington, DC 20208
E-mail: peggy_carr@ed.gov

Chris Chapman
U.S. Department of Education
National Center for Education
Statistics
Room 422K
555 New Jersey Ave., NW
Washington, DC 20208
E-mail: chris_chapman@ed.gov

Lisa Chavez
University of California at Berkeley
16A Forest Street, #41
Cambridge, MA 02138
E-mail: llchavez@uclink4.berkeley.edu

Susan Choy
MPR Associates
2150 Shattuck Avenue
Suite 800
Berkeley, CA 94704
E-mail: schoy@mprinc.com

Arthur L. Coleman
U.S. Department of Education
Office for Civil Rights
330 C Street, SW
Room 5012D
Washington, DC 20202

Joseph Conaty
U.S. Department of Education
National Institute on Student Achievement,
Curriculum, and Assessment
555 New Jersey Ave., NW
Room 510H
Washington, DC 20208
E-mail: joseph_conaty@ed.gov

Lynn Cornett
Southern Regional Education Board
592 10th Street, NW
Atlanta, GA 30318–5790
E-mail: lynn.cornett@sreb.org

James Crouse
University of Delaware
Educational Studies
213H Willard Hall
Newark, DE 19716
E-mail: jcrouse@udel.edu

Nancy Doorey
Delaware State Board of Education
4601 Beechwold Avenue
Wilmington, DE 19803
E-mail: ndoorey@den.k12.de.us

Kimberley Edelin
Frederick D. Patterson Research Institute
UNCF
8260 Willow Oaks Corporate Drive
Fairfax, VA 22031
E-mail: edelink@fdpri.patterson-uncf.org

Ronald Ferguson
John F. Kennedy School of Government
Malcolm Wiener Center for Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail: ron_ferguson@harvard.edu

Michael Feuer
National Research Council
Board on Testing and Assessment
2101 Constitution Avenue, NW
Washington, DC 20418
E-mail: mfeuer@nas.edu

Jeremy Finn
State University of New York
Baldy Hill
Room 408C
Amherst, NY 14260
E-mail: jfinn@acsu.buffalo.edu

Pascal D. Forgione, Jr.
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Suite 500
Washington, DC 20208
E-mail: pascal_forgione@ed.gov

Norm Fruchter
Institute for Education and Social
Policy
New York University
726 Broadway-Fifth Floor
New York, NY 10003

Edward Fuentes
U.S. Department of Education
National Institute on the
Education of
At-Risk Students
555 New Jersey Ave., NW
Room 610E
Washington, DC 20208
E-mail: edward_fuentes@ed.gov

Adam Gamoran
University of Wisconsin
Department of Sociology
1180 Observatory Dr.
Madison, WI 53706
E-mail: gamoran@ssc.wisc.edu

Margaret Goertz
University of Pennsylvania
Consortium for Policy Research
in Education
3440 Market Street
Suite 560
Philadelphia, PA 19104–3325
E-mail: pegg@wfs.gse.upenn.edu

Daniel Goldhaber
The Urban Institute
2100 M Street, NW
Washington, DC 20037

Edmund Gordon
3 Cooper Morris Drive
Pomona, NY 10970

Steven Gorman
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 404G
Washington, DC 20208

David Grissmer
RAND
1333 H Street, NW
Washington, DC 20005–4507
E-mail: david_grissmer@rand.org

Eric Grodsky
University of Wisconsin
Sociology Department
2210 Sommers Avenue
Madison, WI 53704
E-mail: egrodsky@ssc.wisc.edu

Kerry Gruber
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 422B
Washington, DC 20208
E-mail: kerry_gruber@ed.gov

Maureen Hallinan
University of Notre Dame
400 Decio
Notre Dame, IN 46556
E-mail: hallinan.i@nd.edu

Jane Hannaway
The Urban Institute
2100 M Street, NW
Washington, DC 20037
E-mail: jhannawa@ui.urban.org

Robert Hauser
University of Wisconsin
Center for Demography & Ecology
1180 Observatory Drive
Madison, WI 53706
E-mail: hauser@ssc.wisc.edu

Larry V. Hedges
University of Chicago
5835 Kimbark
Chicago, IL 60637
E-mail:
hedge@cicero.spc.uchicago.edu

Lori Diane Hill
University of Chicago
OSC/NORC
1155 East 60th Street
Room 385
Chicago, IL 60637
E-mail: lori@cicero.spc.uchicago.edu

Emily Howard
U.S. Department of Education
Office of the Deputy Secretary
600 Independence Ave., SW
Washington, DC 20202

Christopher Jencks
John F. Kennedy School of Government
Malcolm Wiener Center for Social Policy
79 John F. Kennedy Street
Cambridge, MA 02138
E-mail: jencks@wjh.harvard.edu

Jackie Jenkins
U.S. Department of Education
National Institute on Student
Achievement,
Curriculum, and Assessment
555 New Jersey Ave., NW
Room 611B
Washington, DC 20208

Christopher Johnson
University of Michigan
125 West Hoover Ave.
Apt. 1B
Ann Arbor, MI 48103
E-mail: cjque@umich.edu

Sylvia Johnson
Howard University
2900 Van Ness Street, NW
116 Holy Cross Hall
Washington, DC 20008
E-mail: sjohnson@howard.edu

Will Jordan
The Johns Hopkins University
3505 N. Charles Street
Baltimore, MD 21218
E-mail:
wjjordan@jhunix.hcf.jhu.edu

Dan Kasprzyk
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 422H
Washington, DC 20208
E-mail: daniel_kasprzyk@ed.gov

Andrew Kolstad
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 406B
Washington, DC 20208
E-mail: andrew_kolstad@ed.gov

Daniel Koretz
Center for Testing
Boston College
Campion Hall
Room 3223
Chestnut Hill, MA 02167
E-mail: daniel_koretz@rand.org

Karol Krotki
American Institutes for Research
Education Statistics Services
Institute
1000 Thomas Jefferson St., NW
Suite 400
Washington, DC 20007
E-mail: kkrotki@air-dc.org

Valerie Lee
University of Michigan
School of Education
610 East University Avenue
Room 4220
Ann Arbor, MI 48109–1259
E-mail: velee@umich.edu

Sharon Lewis
Council of Great City Schools
1301 Pennsylvania Avenue, NW
Suite 702
Washington, DC 20004
E-mail: slewis@cgcs.org

Serge Madhere
Howard University
Psychology Department
Washington, DC 20059
E-mail: smadhere@howard.edu

Jo Anne Manswell
Howard University
CRESPAR
2900 Van Ness Street, NW
Holy Cross Hall, Room 427
Washington, DC 20008
E-mail:
jmanswell@crespar.law.howard.edu

John Mazzeo
Educational Testing Service
Rosedale Road
Mail Stop 30-E
Princeton, NJ 08541
E-mail: jmazzeo@ets.org

C. Kent McGuire
U.S. Department of Education
Office of Educational Research
and Improvement
555 New Jersey Ave., NW
Room 600
Washington, DC 20208

Don McLaughlin
American Institutes for Research
P.O. Box 1113
Palo Alto. CA 94302–1113
E-mail: dmclaughlin@air-ca.org

Marilyn McMillen
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 406A
Washington, DC 20208
E-mail:
marilyn_mcmillen@ed.gov

James McPartland
The Johns Hopkins University
CSOS
3003 N. Charles Street
Suite 200
Baltimore, MD 21218

Joseph McTighe
Council for American Private
Education
18016 Mateny Road
Room 140
Germantown, MD 20874
E-mail: cape@connectinc.com

Stephen Morgan
Harvard University
Department of Sociology
William James Hall, 5th Floor
Cambridge, MA 02138
E-mail:
smorgan@wjh.harvard.edu

Ann L. Mullen
U.S. Department of Education
National Center for Education
Statistics Fellow
555 New Jersey Ave., NW
Room 406
Washington, DC 20208
E-mail: ann_mullen@ed.gov

John Mullens
Policy Studies Associates
1718 Connecticut Avenue, NW
Suite 400
Washington, DC 20009
E-mail:
jmullens@policystudies.com

Jennifer O'Day
University of Wisconsin-Madison
Department of Educational Policy
Studies
1000 Bascom Mall
225 Ed. Bldg.
Madison. WI 53706
E-mail:
joday@mail.soe.madison.wisc.edu

Michael Olneck
University of Wisconsin-Madison
School of Education
1000 Bascom Mall
Madison, WI 53706
E-mail:
olneck@mail.soe.madison.wisc.edu

Martin Orland
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 402D
Washington, DC 20208
E-mail: martin_orland@ed.gov

Eugene H. Owen
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 402E
Washington, DC 20208
E-mail: eugene_owen@ed.gov

Jeffrey Owings
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 417A
Washington, DC 20208
E-mail: jeffrey_owings@ed.gov

Audrey Pendleton
U.S. Department of Education
Planning and Evaluation Service
600 Independence Ave., SW
Washington, DC 20208
E-mail:
audrey_pendleton@ed.gov

Robert Petrin
University of Chicago
OSC/NORC
1155 E. 60th Street
Chicago, IL 60637
E-mail:
robp@cicero.spc.uchicago.edu

Gary Phillips
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 408C
Washington, DC 20208
E-mail: gary_phillips@ed.gov

Meredith Phillips
UCLA
School of Public Policy and
Social Research
Public Policy Building
Room 3250
Los Angeles, CA 90098–1656
E-mail: phillips@sppsr.ucla.edu

Paul Planchon
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 413B
Washington, DC 20208
E-mail: paul_planchon@ed.gov

Valena White Plisko
U.S. Department of Education
Planning and Evaluation Service
600 Independence Ave., SW
Room 4143
Washington, DC 20208
E-mail: valena_plisko@ed.gov

John Ralph
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Ave., NW
Room 402L
Washington, DC 20208
E-mail: john_ralph@ed.gov

Stephen Raudenbush
University of Michigan
School of Education
610 East University Avenue
Ann Arbor, MI 48109–1259
E-mail: rauden@umich.edu

Karen Ross
University of Michigan
1072 Island Drive
Apt. 105
Ann Arbor, MI 48105
E-mail: keross@umich.ed

Michael Ross
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 412
Washington, DC 20208
E-mail: michael_ross@ed.gov

Richard Rothstein
Economic Policy Institute
P.O. Box 301
South Welfleet, MA 02663

Laura Salganik
American Institutes for Research
Education Statistics Services Institute
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: lsalganik@air-dc.org

Alex Sedlacek
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Ave., NW
Room 404B
Washington, DC 20208
E-mail: alex_sedlacek@ed.gov

Ramsay Selden
American Institutes for Research
Education Statistics Services Institute
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: rselden@air-dc.org

Sharif Shakrani
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002
E-mail: sharif_shakrani@ed.gov

Becky Smerdon
American Institutes for Research
Pelavin Research Center
1000 Thomas Jefferson Street, NW
Suite 400
Washington, DC 20007
E-mail: bsmerdon@air-dc.org

Marshall S. Smith
U.S. Department of Education
Office of the Deputy Secretary
600 Independence Avenue, SW
Room 6236
Washington, DC 20202

Holly Spurlock
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Avenue, NW
Room 408A
Washington, DC 20208
E-mail: holly_spurlock@ed.gov

Christopher B. Swanson
University of Chicago
OSC/NORC
1155 East 60th Street
Chicago, IL 60637
E-mail: cb-swanson@uchicago.edu

Ricky Takai
U.S. Department of Education
Planning and Evaluation Service
600 Independence Avenue, SW
Room 4111
Washington, DC 20202
E-mail: ricky_takai@ed.gov

Corrine Taylor
University of Wisconsin-Madison
5202 Tolman Terrace
Madison, WI 53711
E-mail: ctaylor@ssc.wisc.edu

Bill Trent
National Academy of Sciences
University of Illinois at Urbana-Champaign
Swanlund Administration Building
601 E. John Street
Champaign, IL 61820
E-mail: w_trent@uiuc.edu

Suzanne Triplett
U.S. Department of Education
National Center for Education Statistics
555 New Jersey Avenue, NW
Room 408D
Washington, DC 20208
E-mail: suzanne_triplett@ed.gov

Roy Truby
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002

Clyde Tucker
U.S. Department of Labor
Bureau of Labor Statistics
2 Massachusetts Avenue, NE
Washington, DC 20212
E-mail: tucker_c@bls.gov

Brenda Turnbull
Policy Studies Associates
1718 Connecticut Avenue, NW
Suite 400
Washington, DC 20009
E-mail:
brenda@policystudies.com

Rob Warren
University of Washington
Department of Sociology
202 Savery Hall
Box 353340
Seattle, WA 98195–3340
E-mail:
jrwarren@u.washington.edu

Walter G. West
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 417B
Washington, DC 20208
E-mail: walter_west@ed.gov

Paul Williams
Educational Testing Service
Rosedale Road
Mail Stop 30-E
Princeton, NJ 08541
E-mail: pwilliams@ets.org

John Wirt
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 402F
Washington, DC 20208
E-mail: jwirt@inet.ed.gov

Alexander Wohl
U.S. Department of Education
Office of the Secretary
600 Independence Avenue, SW
Room 6101
Washington, DC 20202

Shi-Chang Wu
U.S. Department of Education
National Center for Education
Statistics
555 New Jersey Avenue, NW
Room 415
Washington, DC 20208